The Dissertation Committee for Zhifeng Jing Certifies that this is the approved version of the following Dissertation:

Computational Modeling of Protein-ion Binding and Nucleic acids

Committee:

Pengyu Ren, Supervisor

Kevin Dalby

Ron Elber

Hsin-Chih Yeh

Computational Modeling of Protein-ion Binding and Nucleic acids

by

Zhifeng Jing

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

April 2021

Acknowledgements

I would like to thank my advisor, Dr. Pengyu Ren. He has put tremendous effort into this dissertation. Without his acumen and insistence, I would not have successfully started, let alone completed some of the studies. He has also created the best environment for personal growth.

I would also like to thank members of the dissertation committee for their valuable input: Dr. Kevin Dalby, Dr. Ron Elber and Dr. Hsin-Chih Yeh.

I am grateful for colleagues in the Computational Biomolecular Engineering Lab. They have made the projects much easier and fun, although I enjoy the research activities not because they are easy. I appreciate the help from Dr. Changsheng Zhang, Dr. David Bell, Dr. Sara Cheng, Dr. Rui Qi, Dr. Matthew Harger, Dr. Chengwen Liu and Mr. Brandon Walker. Dr. Matthew Harger was a gem in the lab and was instrumental for the development of software that we love and use.

I would like to thank my collaborators for their inspiring ideas and knowledge: Dr. Jay Ponder, Dr. Xu Han, Dr. Jean-Philip Piquemal, Dr. Louis Lagardère, Dr. Marc Thibonnier, Dr. Susan Rempe, Dr. Josh Rackers, Dr. Lawrence Pratt.

Last but certainly not least, I would like to thank my family and friends. My girlfriend Ms. Ming Chen treats me to gourmet cuisine every now and then, which helps distract me from research. My father and mother have always been supportive of my pursuit of doctoral training, even if this means years without family reunion. I feel deeply indebted to them.

Abstract

Computational Modeling of Protein-ion Binding and Nucleic acids

Zhifeng Jing, Ph. D.

The University of Texas at Austin, 2021

Supervisor: Pengyu Ren

Metal ions and nucleic acids are essential for a variety of biological functions. Metal ions play roles in enzyme catalysis, signal transduction and muscle contraction, and the stabilization of protein and nucleic acids structures. Nucleic acids are integral to gene expression and regulation. There are many unsolved questions regarding the function and thermodynamics of metal ions and nucleic acids. Molecular modeling has been an indispensable tool for microscopic understanding of biological processes. While tremendous success has been achieved for the modeling of proteins and organic molecules, it remains challenging to accurate model charged molecules, such as metal ions and nucleic acids. The difficulty mainly arises from the inadequate description of electrostatic interaction and polarization. In this work, accurate models for metal ions and nucleic acids based on AMOEBA polarizable force field were developed. These models along with advanced quantum mechanical methods were then used to study some practical problems. First, the principles underlying Ca^{2+}/Mg^{2+} selectivity in ion-binding proteins were studied. It was shown that the Ca^{2+}/Mg^{2+} selectivity can be explained by many-body polarization, which depends on the chemistry and geometry of the binding

pocket. Second, an existing controversial question regarding the conduction mechanism of potassium channels was resolved by molecular dynamics (MD) simulations with AMOEBA. Contrary to previous beliefs, the conduction operates through nearly ion-saturated states. This mechanism is compatible with almost all existing experimental data. Third, free energy calculation with AMOEBA was used to predict the effect of chemical modifications on the stability of DNA-RNA hybrids, which has implications for the development of gene therapy. Overall, the AMOEBA polarizable force field significantly improves the accuracy for modeling of metal ions and nucleic acids. It is expected that application of polarizable force field will lead to more exciting findings on biological systems.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Note: part of this Chapter is based on a published review paper[1] that I cowrote, Jing et al. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. Annual Review of Biophysics 48, 371-394 (2019).

1. MOLECULAR MODELING AND FORCE FIELDS

Molecular modeling is the use of computer simulations to study the properties of molecules or materials. Molecular modeling bridges microscopic structures and temporal and spatial averaged observables, which is extremely helpful for the understanding and design of new molecules. Molecular modeling is built upon quantum mechanics and statistical mechanics. Quantum mechanics describes the microscopic properties of atoms, while statistical mechanics relates microscopic properties to macroscopic properties. Based on how molecules are represented, molecular modeling can be classified into two categories, ab initio methods and empirical methods. Ab initio methods, also known as first principles methods, does not require any empirical parameters, and therefore can predict the properties of any molecules or materials. However, ab initio methods rely on efficient approximations to solve the Schrodinger equation. Even with the approximations, the computational cost of ab initio methods is very expensive. Typically, the calculations are limited to system size of a few hundred atoms.[2] Current ab initio methods are not accurate enough to consistently predict macroscopic properties of even simple systems such as pure water.[2] Empirical methods, on the other hand, dramatically accelerate the computation by using empirical parameters, at a cost of compromised accuracy and generalizability. The most interesting biological problems, such as protein folding and protein-ligand binding, requires model systems of thousands to millions of

atoms. Therefore, empirical methods are currently the only viable approach for biomolecular modeling. The empirical methods that represent the potential energy of a molecular system by analytical functions are traditionally referred to as molecular mechanics force fields, or simply force fields.[3,4] Other names for this class of methods include molecular models, potential models, and potentials.[5] A comprehensive account of force fields can be found in the reference.[4]

Several sampling methods can be used to calculate ensemble average, including molecular dynamics (MD) simulation and Monte Carlo (MC) simulation. Molecular dynamics (MD) simulation has been a very popular technique for biophysics of proteins and nucleic acids and for drug discovery.[6-8] Even with the efficient force fields, MD still has limited time scale of microseconds for common protein systems, while protein folding and protein-ligand binding take place at a much longer time scale. Thanks to recent advances in hardware and software, MD simulations can now be extended to much larger time and length scales, which is accompanied by the application of MD for many systems. The most notable advances are GPU-accelerated MD simulations[9-11] and enhanced sampling methods.[12-14] These advances allow for, the study of longer timescales and larger length scales using molecular simulations with high precision. At the same time, more problems for existing models are revealed, which requires further model development.

2. POLARIZABLE FORCE FIELDS

Most force fields have separate terms for different energy components: valence interactions, including bond stretching, angle bending and torsion, and non-bonding interactions, including Pauli repulsion, dispersion, and electrostatics. The Pauli repulsion and dispersion terms are also known as van der Waals (vdW). Due to the long-distance

nature of electrostatics, it is both important for biology and computationally intensive. The fixed atomic point charges have been used to represent electrostatics in most force fields because of its simplicity and computational efficiency. However, fixed point-charge force fields also have many limitations.[15-20]

Despite the success of MD simulations with traditional force fields, many challenging systems remain. The challenges have been larger associated with the accuracy of force fields. There have been many studies to improve the accuracy of force fields for MD simulations. The most important areas that need improves are electronic polarization and anisotropic electrostatics interactions. Polarization is the response of the charge distribution to external electric field from its environment. Traditional force field used fixed charges, so polarization is not explicitly represented, but implicitly incorporated into the atomic charges, i.e. the charges are already polarized to be consistent with the aqueous environment. This can be problematic when the properties of interest are related to interfaces or the transfer of substances between environments, such as vacuum, aqueous environment, protein and cell membranes, or when the charge parameters was used for an environment different from the environment that it was designed for. Anisotropic electrostatics interactions arise from the asymmetric charge distributions. In contrast, the point charges in traditional force fields are spherically symmetric. Many biologically relevant molecules have asymmetric charge distributions. The most well-known examples are σ-holes, lone pairs and π-bonding. Methods for incorporating the asymmetric charge distributions have been developed, such as higher-order multipolar electrostatics models [21-25] and/or adding off-center sites.[16,25] It has been shown that the effect of having higher order charge distributions and the effect of polarization are similar in magnitude.[21] This may explain why most polarizable force fields include not only polarization but also anisotropic electrostatics interactions.

Reviews of polarizable force fields can be found in references.[16-19,26-29] The two essential components of polarizable force fields, electrostatics and polarization, are discussed below.

2.1 Electrostatics model

The importance of electrostatic interactions for biomolecules have been the subject of dedicated studies.[17,18] Technically, polarization is also part of the electrostatic interactions. In this section, electrostatics model refers to the model for the nonpolarizable part of electrostatics interactions. The long-range, nonpolarizable part of electrostatic interactions can be well described by the interaction between atomic charges or higher order charge distributions. The parameters in the electrostatics models, i.e. the atomic charges or multipoles, can be determined by charge distributions calculated by QM, or by fitting to electrostatic potential (ESP) from QM calculations or water interaction energy.[30-34] The former approach is common for the determination of atomic multipoles. Molecular charge distributions cannot be uniquely partitioned into atomic contributions. Some useful partition methods include Distributed Multiple Analysis (DMA) and and Iterative Stockholder Analysis (ISA).[35,36]

2.2 Polarization model

Most of the polarization models used in force fields fall in the framework of linear response theory, including induced dipole[26], Drude oscillator[16], and the fluctuation charge model (charge equilibration, or chemical potential equilibration).[37] A less popular approach is modifying the atomic charges based on the geometry through either on-the-fly QM calculation or complicated models. The three popular classical polarization models are discussed below.

Induced dipole model. The induced dipole moment μ^{ind} at each polarizable site can be considered a linear response to external electric field E, $\mu^{\text{ind}} = \alpha E$. The magnitude is related to the atomic polarizability α.[26] The external electric field includes contribution from permanent multipoles and induced dipoles. Therefore, the induced dipole need to be calculated by a linear equation that include all polarizable sites or, in practice, through SCF iterations. The SCF leads to a several times increase in computational cost. More efficient algorithms for induced dipoles have been developed. Because the charge distribution should be smeared, the short-range electric field is damped to calculate the polarization according to the Thole model.[26]

Drude oscillator model.[16] Instead of polarizable sites, the polarization is represented by the movement of Drude particles, which are attached to polarizable sites through springs. The mass of Drude particle is chosen to be smaller than the parent atom to allow for fast response to the electric field, but the mass needs to be sufficiently large to have numerical stability. If the masses of the Drude particles are infinitely small and the positions of the Drude particles are solved to be at ground state, it will have essentially the same computational cost as the SCF induced dipole method. However, the main idea for assigning a fictious mass for the Drude particle is that the Drude particle can be propgated through MD integration, which does not guarantee ground state but significantly improves the computational efficiency. This class of method is called extended Lagrangian integrator. It should be noted that the induced dipole can also be assigned a fictious mass and used with the extended Lagrangian integrator. Similar to the induced dipole model, the short-range electric field is damped to calculate the polarization according to the Thole model.[16] It has been shown that there is no practical difference between induced dipole and the Drude models.[38]

Fluctuating charge (FQ) model. The FQ model is conceptually different from the induced dipole and Drude models. It originates from the principle of the electronegativity equalization. The charges are dynamically determined by the atomic electronegativity, hardness and the external electrostatic potential.[37,39] FQ force fields have been used in simulations of materials and proteins.[37,40,41] The main difference between FQ and the other two polarization models is that the polarization is restricted to between atoms. For example, a planar molecule will not have out-of-plane polarization according to FQ model. This difference is rather trivial, however, as the out-of-plane polarization can be incorporated by virtual atoms connected to the atoms.

Induced dipole and fluctuating charge models have been incorporated into the same force field.[42] However, most modern polarizable force fields use either of the two types of model. There is also no clear distinction between fluctuating charge and polarization. Depending on the specific analysis scheme, the polarization contribution varies from 59.9% to 96.2% for the same molecule.[43]

Considering the similarity between the three polarization models, the choice of polarization model should be decided by the ease of implementation and computational efficiency.

The computational cost of polarizable force field is at least two times that of fixed-charge force field. To allow for treatment of polarization at the cost of fixed-charge force fields, methods for effective polarization have been developed. Several methods for incorporating polarization effect into the atomic charge parameters have been proposed. [32-34] Leontyev and coworkers showed that for simulations of similar configurations, the polarization effect can be described by charge scaling.[44] Charge scaling is equivalent to applying a uniform dielectric constant.[44] Although the charge scaling approach

significantly improves the representation of polarization effect, explicit polarization models are required to achieve better accuracy.[44]

More recent studies have focused on the analysis and improvement of polarization models, specifically the short-range damping model, by QM calculations and energy decomposition analysis (EDA).[45,46] By using QM data on trimers formed by water and ions, the Thole damping model was further improved for distance dependence of polarization.[47]

2.3 Parameterization of polarizable force fields

Because polarizable force fields have a larger number of parameters than fixed-point-charge force field, greater accuracy can be achieved if high-quality data are used in the parameterization. Generally, polarizable force fields have better performance for various thermodynamic state and environments. Polarizable force fields have been successfully applied to gas-phase clusters, ion solvation,[48] and protein-ligand binding.[49] However, some studies found that polarizable force fields could have even worse performance than fixed-charge force fields. This is likely because of the errors in parameters rather than intrinsic problems of polarization.[19,50]

Force field development has been challenging because of the large parameter space and high computational cost.[51] When the same parameterization strategy is used, polarizable force fields will cause more flexibility because of the additional parameters. Fortunately, the parameterization can be designed to be more robust because of the improved physical models. More types of data, including QM EDA are available for the parameterization of PFF.[50,52]

Because of the high computational cost for force field development, human input is necessary for earlier force fields. Recently, automation of parameterization has become

possible thanks to the improvement in hardware and software. Like other model development, force field development requires obtaining reference data and defining an objective function. When a proper guess of initial parameters is available, general optimization algorithm, such as BFGS and least-square optimization is sufficient to produce accurate parameters. For extensive search of the parameter space, more sophisticated optimization algorithms, such as evolutionary algorithms can used.[53,54] Currently, the extensive search is limited to very simple systems such as water and ions.

2.4 AMOEBA force field

AMOEBA (atomic multipole optimized energetics for biomolecular applications), was originally developed by Ren and Ponder and coworkers.[55] Currently, the parameters for metal ions, organic molecules, proteins and nucleic acids are available.[1,26] AMOEBA includes the following energy terms, bond stretching (U_{bond}), angle bending (U_{angle}), torsion (U_{torsion}), bond-angle coupling, bond-torsion coupling, angle-torsion coupling, vdW (U_{vdW}), permanent electrostatics (U_{elst}) and polarization (U_{pol}). Detailed description of the energy terms can be found in the references.[26,56] The bond and angle terms are anharmonic, which is more realistic than popular force fields. For both terms, the anharmonicity is modeled by higher order polynomials and is shared among all bonds/angles. The only adjustable parameters are equilibrium bond/angle values and force constants. The torsion term is modeled as Fourier series up to 6-fold:

$$U_{\text{torsion}} = \sum_{n=1}^{6} \frac{k_n}{2}[1 + \cos(n\phi + \delta_n)] \tag{1.1}$$

where k_n δ_n are the force constant and phase angle parameters, and ϕ is the torsion angle. The phase angle is usually chosen to be $\delta_n = (n\text{-}1)\pi$, but it can also be adjusted when necessary. For proteins and nucleic acids, only up to 3-fold is used. Higher order terms are usually reserved for small molecules.

The vdW term in AMOEBA differs from the traditional LJ-12-6 function,

$$U_{\mathrm{vdW}} = \epsilon_{ij}\left(\frac{1+\delta}{\rho_{ij}+\delta}\right)^{n-m}\left(\frac{1+\gamma}{\rho_{ij}^{m}+\gamma}-2\right), \qquad (1.2)$$

$$n = 14, m = 7, \rho_{ij} = r_{ij}/r_{0,ij}$$

where r_{ij} is the distance, $r_{0,ij}$ is the parameter for equilibrium distance, ε_{ij} is the interaction strength, and δ and γ are the parameters controlling the softness of the interaction. δ and γ are predetermined from fitting to rare gas interactions. The adjustable parameters are ε_{ij} and $r_{0,ij}$. For most vdW interactions between two different atom types, ε_{ij} and $r_{0,ij}$ are determined from the atomic parameters ε_i, ε_j and $r_{0,i}$, $r_{0,j}$ by using arithmetic, geometric or other types of averages.

The permanent electrostatic interactions are modeled by atomic multipoles up to quadrupole. The induced dipoles are determined by the external electric field and the atomic polarizability.

3. FREE ENERGY CALCULATION

One major application of molecular dynamics simulations is binding free energy calculation. The most popular method is the double-decoupling method[57] based on free energy perturbation (FEP) (Figure 1). The double-decoupling method is based on the alchemical transitions that gradually decouple the ligand from its environment during simulations, as if the ligand is in gas phase. The decoupling can be achieved by modifying the parameters for the interaction of the ligand. The double decoupling method

is supported by most MD packages. Two sets of FEP calculations are required, one for the decoupling of the ligand from the binding pocket, and the other from the decoupling of the free ligand from solution. In addition, a free energy correction is needed if the two decoupling simulations are for different standard states. The three free energy changes add up to the binding free energy of the ligand.

The decoupling of the ligand is conventionally parameterized by λ, with λ=0 corresponding to the fully decoupled state and λ=1 corresponding to the full-interaction state. For noncovalent binding, the only interactions between the ligand and the environment are electrostatics (including polarization) and vdW. Sometimes two separate λs are used for electrostatics and vdW. Usually the electrostatics is first decoupled and then vdW is gradually decoupled. Soft-core functions are used for a smooth transition of the vdW interactions.[49] If electrostatics and vdW are parameterized by the same λ, then softcore functions for electrostatics are also needed to avoid slow convergence. This approach in also adopted in some studies[58]

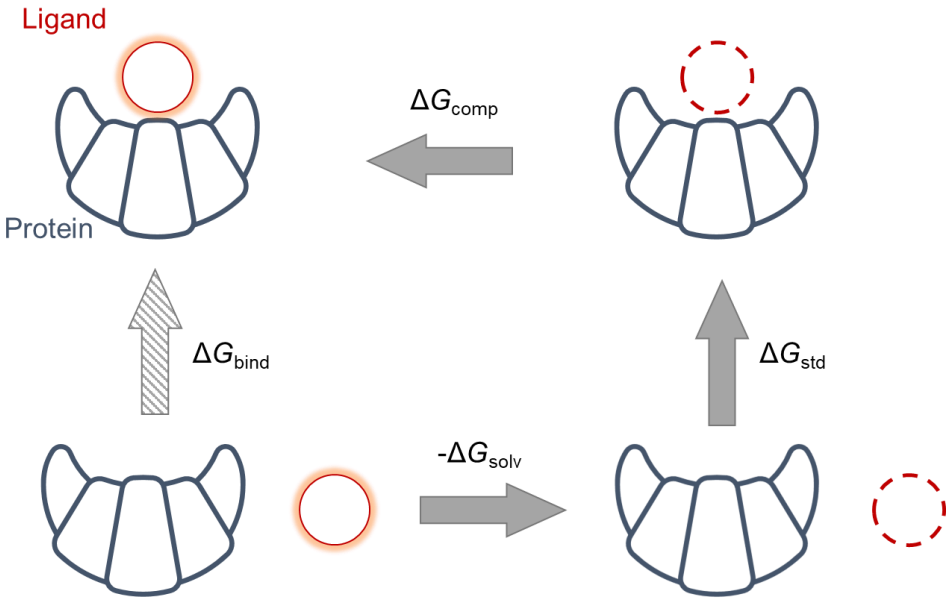


Figure 1: Thermodynamic cycle for calculating binding free energy in the double decoupling method.

Dashed lines indicate decoupled state.

Protein-ligand restraints are usually used in binding free energy calculations so that the ligand does not move away from the binding pocket even when it is decoupled from the system. The free energy change ΔG_{std} from the standard state in gas phase to the restrained state in gas phase can be evaluated either analytically or numerically. It should be positive since the restrained volume is usually very small. Two types of restraints have been proposed: (1) distance-based restraint with either harmonic or flat-bottom restraint function; and (2) Boresch restraints, where a total of 6 distance/angle/torsion restraints limit the relative transition and rotation of the ligand with respect to the protein.[57] For the Boresch restraint, the free energy correction is calculated by

$$\Delta G_{\text{std}} = RT \ln \left[\frac{8\pi^2 V \left(k_r k_{\theta_A} k_{\theta_B} k_{\phi_A} k_{\phi_B} k_{\phi_C}\right)^{\frac{1}{2}}}{r_{aA,0}^2 \sin\theta_{A,0} \sin\theta_{B,0} (2\pi RT)^3} \right] \tag{1.3}$$

For more complicated types of restraints, the free energy correction can be calculated by numerical evaluation.

Chapter 2: AMOEBA+ force field for metal ions

1. INTRODUCTION

A major challenge of force fields is the transferability or robustness for different systems and thermodynamic conditions. Force fields are approximate representations of the true potential energy surface that allow for large-scale simulations. Therefore, the goal of force fields-based simulations is not to predict accurately the entire potential energy surface, but to predict structural and thermodynamic properties of interest. Force field development is based on limited QM and experimental data. Through careful parameterization, reasonable accuracy can be achieved for the fitting data sets. However, the accuracy for the fitting data set does not always transfer to the accuracy for real systems.

One way to improve the transferability of force fields is to use more physically grounded functional forms. Formally, this is a way of regularization that can significantly improve the quality of models while requiring relatively small data sets. Polarizable force fields, such as AMOEBA, improves on fixed-charge force field by including explicit treatment of polarization. Polarization varies significantly between environment, such as gas phase, aqueous solution, ligand binding sites, lipid membranes. In fixed charge force fields, polarization is absorbed into the partial charges, so the transferability between different environment is poor.

Besides polarization, some other energy components are not well captured by traditional force fields. For example, electrostatic interactions are represented by point charges in traditional force field. Although atoms can be treated as point charges at a distance, this approximation does not work very well at short range. The electron cloud

around each atom is not perfectly symmetrical as modelled by a point charge, and the interaction between electron clouds are much weaker than those between point charges due to the so-called charge-penetration effect. In addition, the charge distribution depends on the geometry of molecules, such as bond length and angles. When molecules move between different phases, there could be nonnegligible changes in the bond lengths and angles. Another example is charge transfer effect, which is quite different from electrostatic interactions or polarization in classical models.

AMOEBA+ is an extension of AMOEBA to include more robust representations of polarization,[59] charge penetration,[60,61] charge transfer[62] and geometry-dependent charge-flux.[63,64] Previously, AMOEBA+ was shown to have superior accuracy for the geometry and energies of water in gas phase while having similar accuracy for liquid properties compared to AMOEBA, which indicates the good transferability of AMOEBA+. In this work, the AMOEBA+ parameters for metal ions were developed to reproduce a variety of properties.

2. METHOD

In AMOEBA+, nine parameters are required to describe an atomic ion: vdW (radius and well depth), polarization (polarizability, damping for direct polarization, damping for mutual polarization, Eq.(2.2)), charge transfer (prefactor and exponent, Eq. (2.3)), and charge penetration (effective charge and exponent, Eq. (2.1)).

$$V_q = [Z_{\text{eff}} - (Z_{\text{eff}} - q)(1 - \exp(-\alpha_{\text{CP}} r))]/r \tag{2.1}$$

$$\mu_i^{\text{ind}} = \alpha_i \left(\sum_j T_{ij}^{\text{damped}} M_j + \sum_j T_{ij}^{\text{damped}} \mu_j^{\text{ind}} \right) \tag{2.2}$$

$$f_{\text{Thole}}(r) = 1 - \mathrm{e}^{-au(r)^3}$$

$$f_{\mathrm{MB}}^{\mathrm{direct}}(r) = 1 - \mathrm{e}^{-a^{*}u(r)^{3/2}}$$

$$E_{\mathrm{CT},ij} = -A_{\mathrm{ij}} \exp\left(-B_{ij} r_{ij}\right) \quad (2.3)$$

For fixed charge force fields, including AMBER, CHARMM and OPLS, the adjustable parameters are vdW radius and well depth. For AMOEBA, the adjustable parameters are vdW radius and well depth, atomic polarizability and polarization damping. Ideally, each term should be fitted separately, but these energy components are not observables and there are no methods for stable and meaningful calculations of these energy components. In particular, the decomposition of charge transfer is an open question.[65,66] Good transferability can be achieved if a force field can accurately reproduce all energy components. However, if there are errors in energy components, the errors could accumulate and the error in total energy could be significant. In this case, emphasis should be placed on reproducing the total energy.

To determine the parameters for all components, different types of data were used, including the interaction energy and its components for ion and ion-water complexes, (Figure 2, Table 1), and experimental properties of salt solutions and ionic crystals.

The equilibrium structures for ion-water dimers were obtained from previous work[67] or QM geometry optimization. Unless otherwise noted, all geometry optimizations were performed using MP2 combined with the aug-cc-pVTZ basis set. Then a rigid distance scan was performed to obtain structures with ion-water distance from 0.8 to 3.0 times the optimized value. CCSD(T)/CBS energy was calculated as MP2/CBS plus the higher order perturbation δCCSD(T) using the aug-cc-pVTZ basis set. SAPT/aug-cc-pVQZ was used for the energy decomposition analysis.

The structures for ion-water clusters, which include one ion and 2 to 6 water molecules, were optimized by using MN15/aug-cc-pVDZ followed by MP2/aug-cc-pVTZ, with initial structures from MD simulations at 298 K using AMOEBA. Gaussian16, ORCA 4 and Psi4 were used for the geometry optimization and energy calculations. The interaction energy between the ion and all the the water molecules was calculated by using MP2/CBS with a two-point extrapolation from MP2/aug-cc-pVTZ and MP2/aug-cc-pVQZ.

Ion dimers in the reference data include both homo-dimers and hetero-dimers. The equilibrium structure for hetero-dimers were optimized by QM, while the ion-ion distance in the reference structure for homo-dimers was set to the equilibrium ion-water distance. Then as series of structures was generated by varying the ion-ion distance. CCSD(T)/CBS energy was calculated as MP2/CBS plus the higher order perturbation. SAPT/aug-cc-pVQZ was used for the energy decomposition analysis.

The linear ion trimers consist of one anion at the center and two cations at symmetric positions. The optimal ion-ion distance was optimized by QM with the symmetry imposed. The target data is total energy minus polarization energy. The QM total energy and polarization energy were calculated by CCSD(T) and ALMO-EDA2//ωB97X-V/def2-QZVPPD, respectively.

The lattice energy and lattice constant of the salt crystals were also included. To allow for more consistent unit in the data set, the lattice energy and constant were not used directly in the fitting, but rather to create an energy curve by scaling the AMOEBA energy curve according to the lattice energy and constant. The curve contains five points, -0.5, -0.25, 0, 0.25, 0.5 Å plus the equilibrium lattice constant.

The activity coefficient γ of salt solution was included as a validation. γ is related to ionic hydration free energy,[68] as shown in Eq. (2.4)

$$\beta \Delta A_{\mathrm{solv}} = 2 \ln \gamma_{\pm} + 2 \ln \left[\frac{\rho_{\mathrm{w}}}{\rho_{\mathrm{b}}} \left(1 + \frac{N_{\mathrm{s}} M_{\mathrm{s}}}{N_{\mathrm{w}} M_{\mathrm{w}}} \right) \right] \qquad (2.4)$$

where ρ_b and ρ_w are the densities. *N* and *M* are the number and molar mass. "b", "w" and "s" denote solution with a molarity *b*, water and salt, respectively.

Name	Component	Properties
Ion-water dimer	M + H_2O or X + H_2O	Interaction energy Energy decomposition
Ion-water cluster	M + $(H_2O)_n$ or X + $(H_2O)_n$, $2 \leq n \leq 6$	Interaction energy
Ion dimer	M-X, M-M, X-X	Interaction energy Energy decomposition
Linear ion trimer	M-X-M	Energy decomposition
Crystal	Salts containing Na and/or Cl	Lattice energy/length

Table 1: Reference data used in the parametrization.

M and X denotes alkali metal and halogen ions.

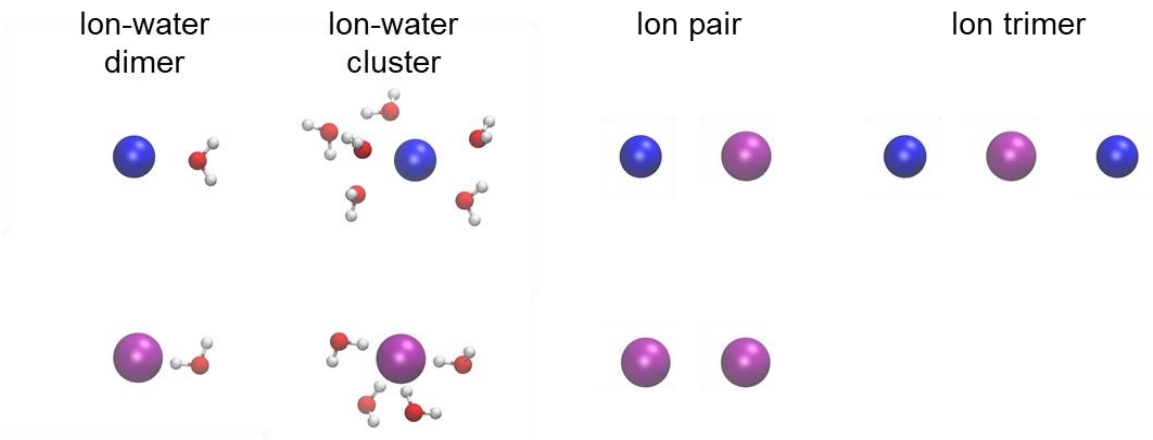


Figure 2: Examples of ion complexes used in QM calculations.

Blue and purple balls represent alkali metal and halogen ions, respectively; red/white balls represent oxygen/hydrogen atoms.

Parameter optimization. Some of the reference data points involve two different ions, so the ion parameters are correlated with each other. Therefore, the parameters for each ion cannot be fitted separately. One possible approach is to fit all ion parameters at the same time. However, this approach is very time consuming, and the final parameters may not be optimal. In this work, a two-step approach was used. In the first step, Na+/K+/Cl- parameters were optimized at the same time. After Na+/K+/Cl- parameters were determined, remaining ions were optimized independently by using interactions with water and Na+/Cl-. The effective charge parameters of Br- and I- were set to the effective nuclear charge determined from previous work.[69] There is no need to adjust the effective charge parameters for other ions. The objective function is not very sensitive to atomic polarizability parameters, so they were fixed at the values of the respective atoms. In the first step, parameter scan for certain force field terms (polarization damping and charge-transfer *B*) were used to find the best solution. In the second step, the scan was performed on damping parameter. This two-step parameter scan approach reduces the number of adjustable parameters to 4/5: r_0, ε, *A* (and *B*), and α_{CP}. The objective function consists of squared errors and a regularization term,

$$L(\mathbf{x}) = \sum_i w_i (A_i(\mathbf{x}) - A_i^0)^2 + \alpha \sum_j \left(\frac{x_j - x_j^0}{x_j^0} \right)^2 \qquad (2.5)$$

where A_i^0 is a specific property in the reference data, A_i is value calculated by force field, w_i is the weight, x_j^0 and x_j are the initial and current parameter, and α is the strength of the regularization term. The parameters in the equation other than **x** are also known as hyperparameters. The Trust Region Reflective algorithm was used for optimization of the objective function. The optimization was performed by using python with SciPy package.

The weights w_i were chosen according to the error and relevance of the data points. For convenience, the total weight is described as the product of a relative weight between different types of data and a weight within each type of data (Table 2 and Table 3). Ion-water interactions were considered more important than ion-ion interactions, the total energy was considered more important than energy components and therefore assigned 8x the weight, and among the components electrostatics was considered the most reliable and therefore assigned 4x the weight. Furthermore, the 3 data points with lowest energy were important for the prediction of equilibrium interaction distance so there were assigned 100x the weights. The hydration free energy is more reliable than the lattice energies, so it was assigned much larger weight. The relative weight with one type of data is inversely proportional to the squared error, whereas the error is calculated by the difference between larger and smaller basis sets with a minimum of 0.1 kcal/mol:

$$\sigma = \max\{0.1, |E_{\mathrm{aTZ}} - E_{\mathrm{ref}}|\} \ (\mathrm{kcal/mol}) \qquad (2.6)$$

The calculated errors are 0.3~1.0 kcal/mol for equilibrium interaction energy of ion-water dimers and 1~3 kcal/mol for equilibrium interaction energy ion dimers and trimers. Since various types of data were used in the optimization, a small regularization factor $\alpha = 1.5 \times (n_{\mathrm{data}}/n_{\mathrm{parameter}})^2$ was enough to prevent unphysical force field parameters.$\alpha = 1.5 \times (n_{\mathrm{data}}/n_{\mathrm{parameter}})^2$.

Name	Relative Weight
Ion-water dimer	1
Ion homodimer	10^{-2}
Ion heterodimer	10^{-2}
Linear ion trimer	10^{-6}
Ion-water cluster	1

Table 2. Relative weight between different types of QM data.

Data type	Weight
Dimer/trimer interaction energy, 3 lowest points	$8\times10^{-4} \times (\sigma/\text{kcal}\cdot\text{mol}^{-1})^{-2}$
Dimer/trimer interaction energy, other points	$8\times10^{-2} \times (\sigma/\text{kcal}\cdot\text{mol}^{-1})^{-2}$
Dimer electrostatics energy	$4\times10^{-2} \times (\sigma/\text{kcal}\cdot\text{mol}^{-1})^{-2}$
Dimer vdW/induction energy	$1\times10^{-2} \times (\sigma/\text{kcal}\cdot\text{mol}^{-1})^{-2}$
Ion-water cluster interaction energy	$8\times10^{-2} \times (\sigma/\text{kcal}\cdot\text{mol}^{-1})^{-2}$
Lattice energy	4
Hydration free energy	10^{4}

Table 3. Weights accounting for the error and importance of reference data.

3. RESULTS

3.1 Gas-phase interactions

The interaction energies calculated by the optimized parameters are compared to QM in Figure 3, Figure 4. The interaction energies for alkali metal ion-water dimers

except for Li+ were predicted by AMOEBA+ with reasonable accuracy. For Li+, equilibrium distance is underestimated while the equilibrium interaction strength is overestimated. Examination of the energy components shows a large error in the induction energy, which was not captured by other components. The total energy of alkali metal ion homodimers was accurately reproduced. For Na+ homodimer, the short-range repulsion was underestimated by AMOEBA+, which can be attributed to the error in induction energy. The AMOEBA+ functional form is flexible enough to reproduce both ion-water and ion homodimer interactions for N+, and these errors are caused by the requirement of experimental hydration free energies. The interaction energies of ion-water clusters were predicted with good accuracy for smaller ions.

The interaction energy of halogen ion-water dimer was also reasonably reproduced by AMOEBA+ (Figure 4). In particular, both total energy and components were accuracy reproduced except for F-, in which case total energy is accuracy but the energy components have relatively large errors. For the halogen ion homodimers, the short-range interaction is attractive, which seems counterintuitive. The electrostatic energy of Cl/Br/I homodimers becomes negative at close separation. The total energy and components were by AMOEBA+ except for F-, which shows an error cancelation between electrostatics and vdW. The negative electrostatics energy in halogen homodimers anions is due to charge penetration. Without charge penetration, such as in AMBER and AMOEBA, the electrostatics is modeled as point charges/multipoles, so the electrostatic energy is greater than zero. The importance of charge penetration in halogen homodimers is also the reason why it is included in the reference data. Preliminary data showed that found that without the homodimer data, the charge-transfer parameters will be ill-determined. The interaction energy of ion-water clusters was reproduced for Cl-/Br-

. The accuracy for F-/I- ion-water clusters was sacrificed to better reproduce the hydration free energies.

The trends for ion heterodimers were similar to those of ion-water dimers (Figure 5). Generally, interactions with Li+/F- have relatively larger errors due to their covalent-bonding tendency. AMOEBA+ underestimated the equilibrium distance and overestimated the interaction strength for Li-Cl and Na-F.

Linear ion trimers have similar performance compared to the ion heterodimers (Figure 6). Since SAPT only support dimers, the energy decomposition was calculated by ALMO-EDA, and the total energy minus polarization energy was compared to AMOEBA+ total energy. Since the polarization in AMOEBA+ is only represented by induced dipoles, there is no polarization on the central halogen ion, while the polarization on the two cations is negligible because of their small polarizability. For the trimers without Li+ or F-, the error in the equilibrium interaction energy for the trimers excluding Li+/F- is less than 5%, and the error in equilibrium distance is around 0.1 Å. The interaction strength for Li+ and F- are stronger in AMOEBA+ compared to the reference QM.

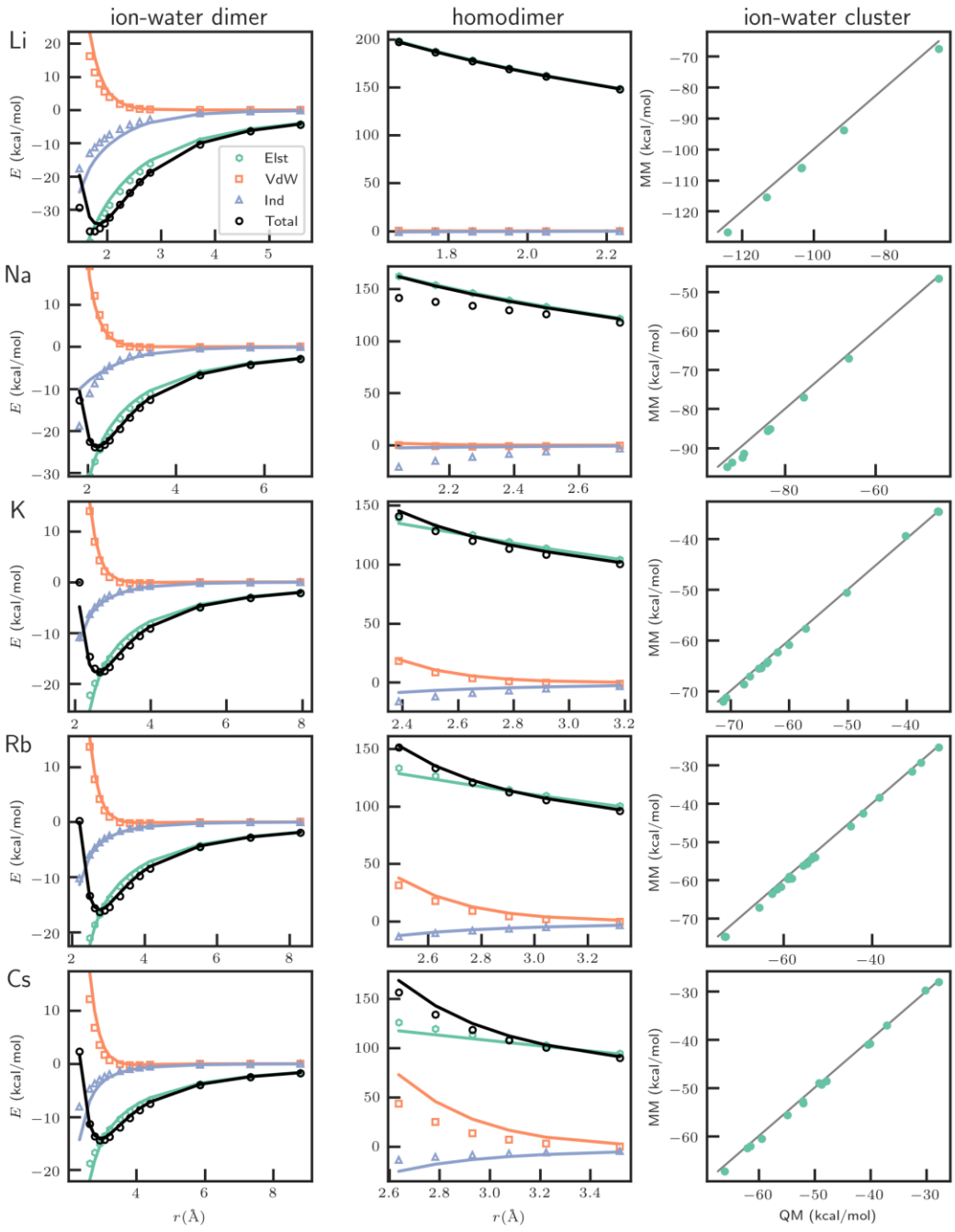


Figure 3: Ion-water and ion homodimer interactions for alkali metal ions.

Induction is defined as polarization plus charge transfer in AMOEBA+. For the dimer interactions, QM and MM are represented as lines and symbols, respectively.

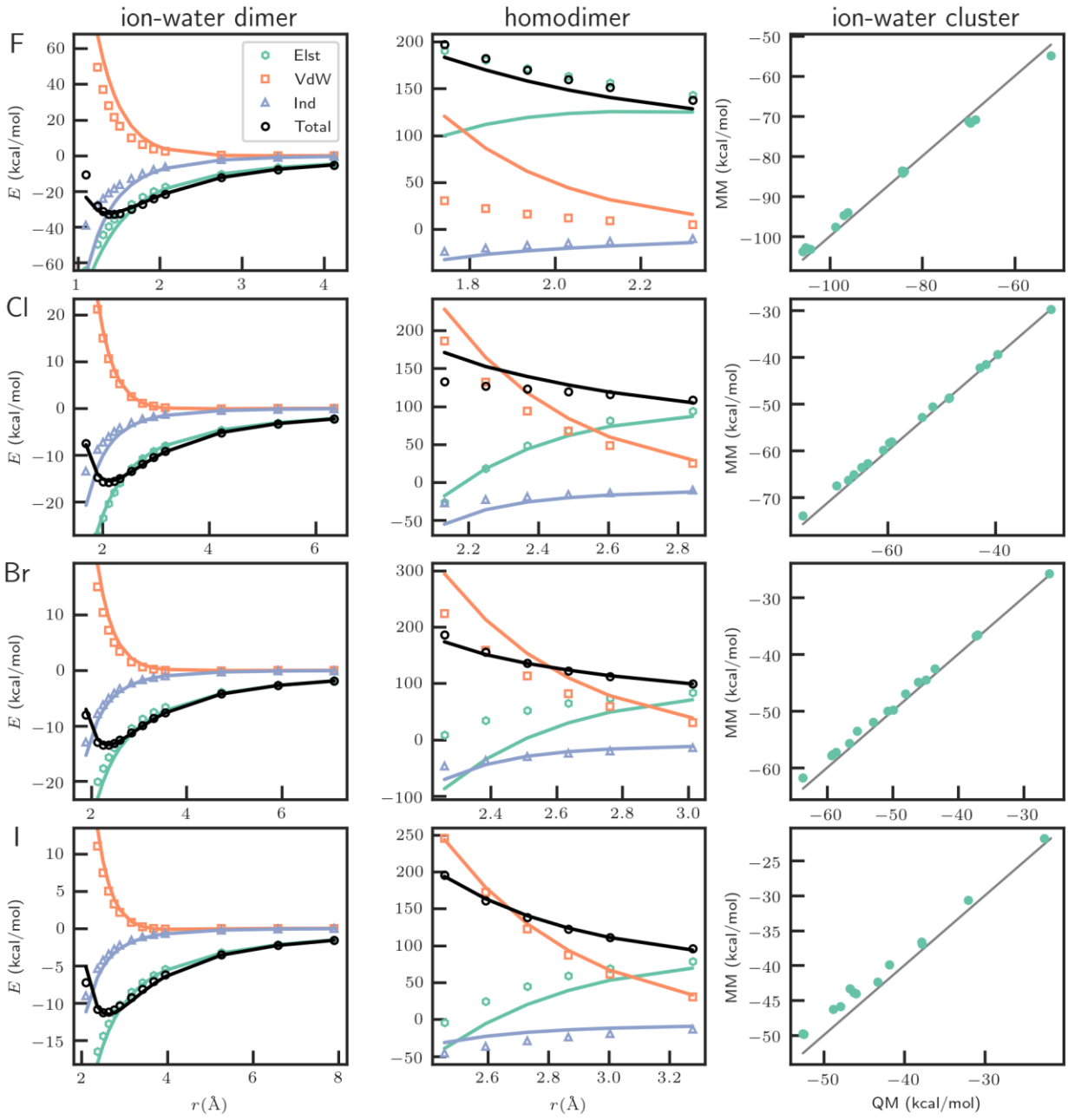


Figure 4: Ion-water and ion homodimer interactions for halogen ions.

See
Figure 3 caption for description.

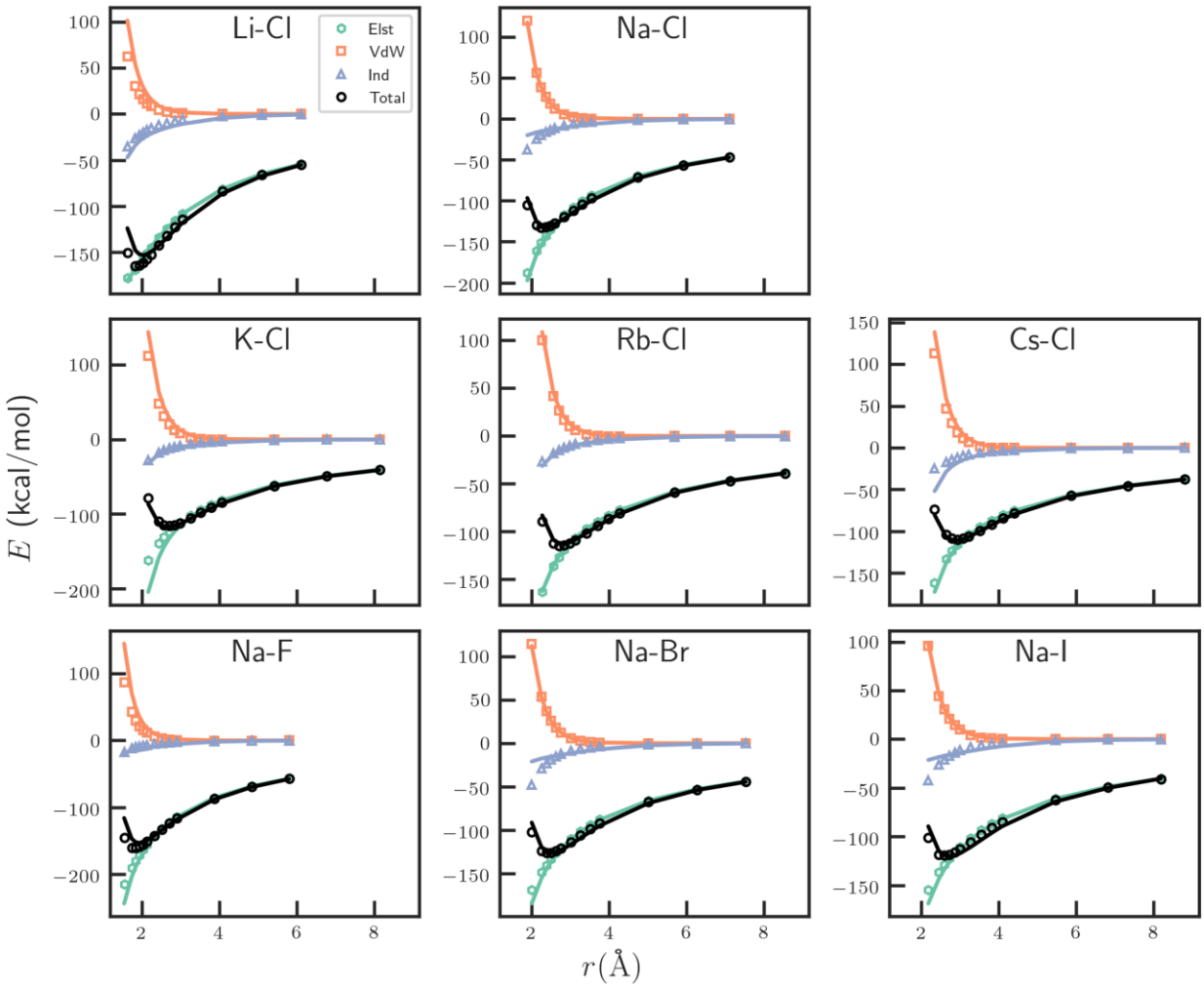


Figure 5: Interaction energies and components for ion heterodimers.

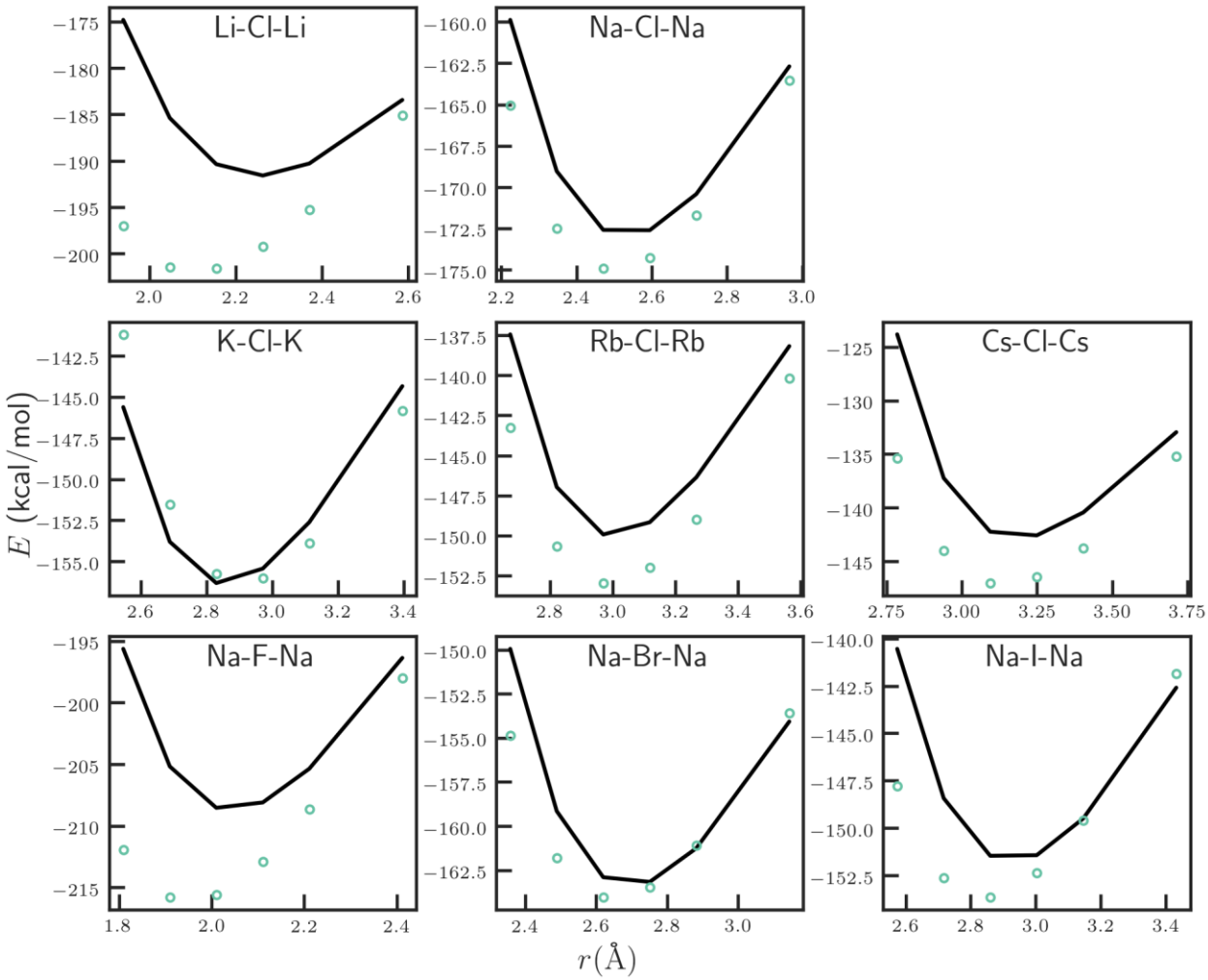


Figure 6: Interaction energies with polarization excluded for linear ion trimers.

QM and MM energies are represented as lines and symbols. The QM energy is calculated by CCSD(T)/CBS interaction energy minus the polarization energy from ALMO-EDA.

3.2 Lattice energies

The lattice energies of crystals containing Na and/or Cl were used in the parameterization, while those of other crystals were used as validation (Table 4). Most of the lattice constants were reproduced within 0.2 Å, except for LiF, KCl and KBr which

have relatively large errors in lattice constants. LiF, CsBr and CsI have relatively large errors in lattice energy. Overall, the accuracy is similar as Wang.[68] Improvements in lattice energies were found for crystals containing Li and/or F. Some of the lattice constants were also improved. Poorer performance for some other crystals were also identified. However, the AMOEBA+ model should be more transferable because it does not use special pairwise vdW parameters.

	Lattice constant *a* (Å)			Lattice energy (kcal/mol)		
	Expt.	AMOEBA+	Deviation	Expt.	AMOEBA+	Deviation
LiF	3.99	3.73	-0.26	-250.7	-266.0	-15.3
LiCl	5.08	5.02	-0.05	-206.5	-207.3	-0.8
LiBr	5.43	5.39	-0.03	-196.0	-195.7	0.3
LiI	5.90	5.85	-0.06	-182.6	-183.3	-0.7
NaF	4.59	4.61	0.02	-222.3	-219.1	3.2
NaCl	5.58	5.72	0.14	-188.8	-187.7	1.1
NaBr	5.91	6.06	0.15	-180.2	-179.5	0.7
NaI	6.39	6.55	0.16	-168.5	-168.0	0.5
KF	5.30	5.42	0.13	-198.1	-196.7	1.4
KCl	6.23	6.44	0.20	-172.1	-171.6	0.5
KBr	6.52	6.75	0.22	-165.2	-163.8	1.4
KI	6.98	7.15	0.17	-155.4	-155.4	0.0
RbF	5.58	5.57	-0.01	-190.0	-190.2	-0.2
RbCl	6.52	6.68	0.16	-166.1	-167.7	-1.6
RbBr	6.82	6.99	0.17	-159.7	-162.6	-2.9
RbI	7.26	7.24	-0.01	-151.1	-154.6	-3.5
CsF	5.96	5.86	-0.10	-181.4	-182.9	-1.5
CsCl	4.07	4.17	0.10	-160.1	-162.3	-2.2
CsBr	4.24	4.20	-0.03	-154.6	-160.7	-6.1
CsI	4.51	4.45	-0.06	-146.5	-152.5	-6.0

Table 4: Lattice constants and lattice energies for salt crystals.

3.3 Hydration free energies

The ion hydration free energies from Schmid were used as part of the target for parametrization. Previous AMOEBA ion parameters[48] were also consistent with the Schmid values. Other than Li+, hydration free energies for alkali metal and halogen ions were reproduced within the error of experiments (Table 5). The relatively larger error for Li+ is due to comprise between QM and experimental data. If QM energies for Li+ are accuracy reproduced, the predicted hydration free energy will be much smaller than the experimental value. Therefore, the different types of data for Li cannot be reproduced at the same time.

	Experiment			AMOEBA+	dev.
	Schmid[70]	Marcus[71]	Tissandier[72]		
F-	-119.74	-114.96	-104.39	-120.19	-0.45
Cl-	-89.15	-85.09	-74.61	-89.35	-0.20
Br-	-82.70	-79.11	-68.20	-82.95	-0.25
I-	-74.33	-69.55	-59.26	-74.48	-0.15
Li+	-113.77	-117.35	-128.43	-111.37	2.40
Na+	-88.67	-91.06	-103.19	-88.37	0.30
K+	-71.22	-74.33	-86.03	-71.15	0.07
Rb+	-65.97	-69.55	-80.60	-65.68	0.29
Cs+	-60.00	-63.58		-59.72	0.28

Table 5: Ion hydration free energies (kcal/mol) from experiments and AMOEBA+.

The statistical error for AMOEBA+ is around 0.3 kcal/mol. Schmid results are used as the reference data.

3.4 Ionic activity coefficients

The ionic activity coefficients measure the aggregation of ions in aqueous solution. As shown in (Figure 7, Figure 8), the activity coefficients for NaCl and KCl

were reasonably reproduced. Even at high concentration, there is no significant deviation from the experimental values. Previously, it was found that AMOEBA and fixed-charge force fields lead to spurious aggregation of ions at high concentration.[68] The accuracy of AMOEBA+ in the ionic activity coefficients guarantees that AMOEBA+ is free from such a problem for simulation of salt solutions.

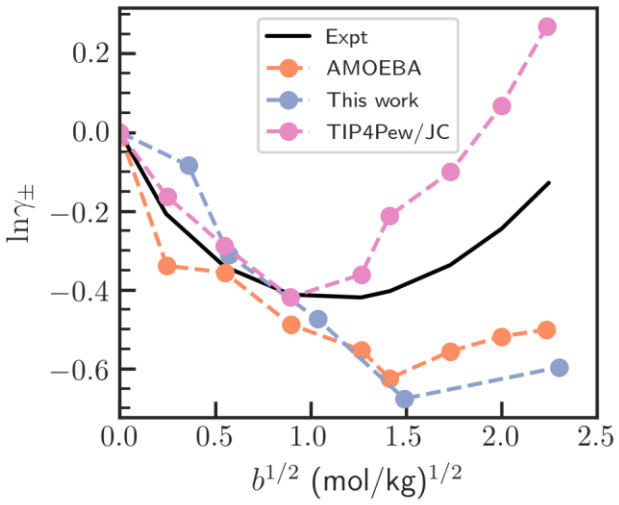


Figure 7: Mean ionic activity coefficient in NaCl solution.

All data except for “This work” are taken from Wang.[68]

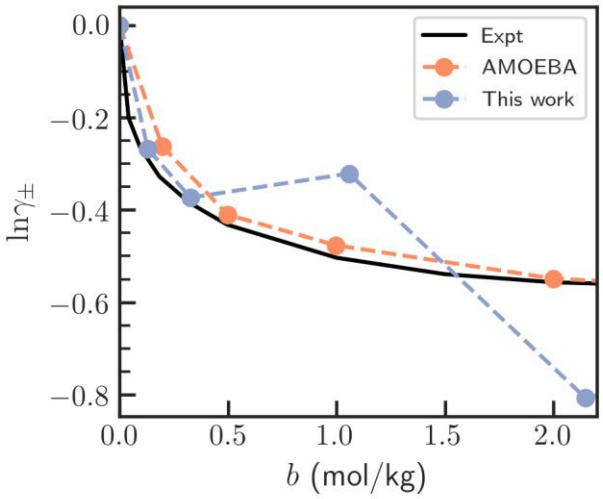


Figure 8: Mean ionic activity coefficient in KCl solution.

4. SUMMARY

The AMOEBA+ parameters for main-group monovalent ions was developed by using QM interaction energy, energy decomposition and condensed-phase experimental data. Various types of data were included, such as ion-water dimer, ion-water cluster, ion homodimers and heterodimers, ion trimers. These types of data are necessary for transferability of the final parameters. Generally, both the total energy and energy components are well reproduced by AMOEBA+, which is mainly attributed to the improved charge-penetration and charge-transfer model. In particular, the negative electrostatics energy in halogen homodimers was reproduced with the help of the charge-penetration model.

In the original AMOEBA ion parameters, the hydration free energy was reasonably predicted by using only ion-water dimer data. However, the AMOEBA parameters have large errors for the interactions between ions. If only ion-water dimer data was used to develop AMOEBA+, the resulting parameters will predict a range of

hydration free energy values. Therefore hydration free energy needs to be included in the reference data. Compared to AMOEBA, AMOEBA+ can reproduce a variety of data include both ion-water and ion-ion interactions.

The AMOEBA+ parameters developed in this work will be useful for the simulations of gas-phase ion-water clusters as well as aqueous solutions and protein-ion binding.

Chapter 3: Theoretical study of protein-ion interactions

Note: this Chapter is based on a published paper[73] for which I conducted the experiment, analyzed the data and wrote the paper, Jing, Z., Qi, R., Liu, C. & Ren, P. Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field. J. Chem. Phys. 147, 161733 (2017).

1. INTRODUCTION

Metal ions are essential for a variety of biological functions, including enzyme catalysis, signal transduction and muscle contraction, and the stabilization of protein and NA structures.[74] A substantial fraction of enzymes have metal centers in the active sites.[75] Metal ions can either directly participate in redox reactions or stabilize reaction intermediates and transition states.[75] Metal ions such as Ca^{2+} and Zn^{2+} are versatile messengers for cell signaling.[76,77] Ca^{2+} signaling is involved in various processes including metabolism, secretion, cell proliferation and cell death.[77,78] Like many other metals, Ca^{2+} at elevated levels can be cytotoxic. Thus precise control of Ca^{2+} is important.[77,78] Mg^{2+} is a structural component of tertiary structures of RNA and DNA. It is also required for the activity of DNA/RNA polymerases.[79]

Many diseases are related to mutations at metal-ion binding sites, and over 1/8 of disease-associated single nucleotide polymorphisms (dSNP) are associated with first- and second-shell residues around metal ions.[3] Also, abnormal neuronal calcium signaling has been observed in many neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease.[4] A large portion of currently used drugs target ion-binding proteins, and those targeting ion-channel proteins alone contribute to ~15%. However, most of the

ion-channel proteins remain unexplored for therapy, mainly due to our poor understanding of the roles of ion channels.[5] Metallodrugs that target DNA G-quadruplex have been widely utilized for chemotherapy, and there is a need to reduce their serious side effects.

Molecular modeling has been an indispensable tool for microscopic understanding of biological processes and the development of novel therapeutic compounds. While the models for proteins and organic molecules have been developed and extensively validated over the past decades, metal ion modeling remains difficult. Compared to organic molecules, metal ions have strong electrostatics interaction, polarization, and other quantum effects. These complicated interactions require high accuracy of the force field. Computational efficiency or adequate sampling is another challenge for molecular modeling. There is a gap between timescales achievable with common simulation methods and time scales relevant for biological functions. Proteins can adopt conformational changes upon binding to metal ions. These processes take at least hundreds of microseconds,[14] compared to the timescale of tens of nanoseconds accessible to polarizable molecular dynamics simulations. Recent advances in sampling methods,[15] algorithms and computer hardware[80] have provided the opportunity to reach biologically relevant timescales. However, systems with highly charge species, such as metal ions, remain challenging for modeling. In charge systems, there are complicated short-range electrostatics interactions, charge transfer and polarization. These different types of interactions come together to shape the specificity and selectivity of protein-ion binding.[81,82] Popular biomolecular force fields, such as AMBER,[83,84] CHARMM, [85] OPLS-AA,[86] and GROMOS[87], do not provide accurate representations of these effects.

By improving the electrostatics and polarization models, polarizable force fields are promising for accurate modeling of metal ions.[27,88-93] For example, AMOEBA has very good accuracy for both gas-phase ion-water clusters and metalloproteins.[94-96] The charge transfer, which may also be significant for metal ion interactions, [97] have been less well discussed for PFFs. The charge-transfer term is only included in a small number of PFFs, such as SIBFA (sum of interactions between fragments ab initio computed).[98]

Protein-ion interactions can be quatititively understood by using ab initio quantum mechanical (QM) calculations. Because of the high computational cost of QM, small reduced model compounds are often used.[99] One recent development of QM methods is energy decomposition analysis method (EDA), which provides rich information for various types of interaction. Symmetry adapted perturbation theory (SAPT) is one major EDA method which decompose the total interaction energy into contributions from electrostatic, exchange-repulsion, dispersion and induction.[99] SAPT has been useful for both the understanding of molecular interactions and the development of next-generation force fields.

In this work, EDA methods are used to understand various components in protein-ion interactions, especially many-body energy and charge transfer. This understanding will also be helpful for the evaluations of traditional non-polarizable force fields and polarizable force fields.

2. METHODS

2.1 AMOEBA

The functional forms of AMOEBA have been described in previous review papers.[26] The adjustable parameters for Mg and Ca ions include vdW (well depth and radius) and polarization parameters (atomic polarizability α and damping parameter a).

The polarization damping parameter softens the electric field at short range according to a smeared charge distribution

$$\rho = \frac{3a}{4\pi}\exp(-au^3) \tag{3.1}$$

where a controls the length scale of the smeared charge distribution, and $u = R_{ij}\left(\alpha_i\alpha_j\right)^{-1/6}$ is a normalized distance between two atoms i and j and their polarizabilities. In AMOEBA, $a = 0.39$ is shared among all organic molecules, water and monovalent ions. For divalent metal ions, it may be necessary to use different a values.[73]

2.2 Energy decomposition analysis

Energy decomposition analysis is a theoretical framework for the analysis of molecular interactions. The interactions energies are decomposed or separated into physically meaningful terms. contributions. This is achieved through either perturbation or restraints on the electronic structure. ALMO-EDA (absolutely localized molecular orbitals)[65] decomposes the total interaction energy into frozen orbital interaction $E_{\mathbf{frz}}$, polarization energy $E_{\mathbf{pol}}$, and charge-transfer (CT) term. Symmetry adapted perturbation theory (SAPT) is based on perturbation theory, and different energy components including electrostatic, exchange-repulsion, dispersion and induction are calculated separately and they sum up to be the total interaction energy.[100,101]

2.3 Interaction energy and many-body expansion

The total interaction energy is defined as the difference between total energy and the sum of energies of isolated fragments. This is expressed as $E_{int} = E_{total} - \sum_i E_i$. The definition of fragments can be arbitrary depending on the quantities of interest. Usually, the fragments only include whole molecular fragments. Here, the fragment

contains one or more molecules. The fragment can be either one ligand or multiple ligands. The many-body energy is defined as calculated as difference between total energy and the sum of two-body energies, and two-body energies is defined as the interaction energy in an isolated dimer.

$$E_{MB} = E_{int} - \sum_{i,j} E_{int}(i,j) = E_{total} - \sum_i E_i - \sum_{i,j}\left(E_{total}(i,j) - E_i - E_j\right), \quad (3.2)$$

where subscript "MB" denotes many-body energy and "int" denotes interaction energy.[73]

2.4 Computational details

The most common ligands in Mg/Ca binding sites are water, Asp/Glu/His side changes and protein backbone carbonyl groups. To mimic these interactions, water, acetate, acetamide and imidazole were used. Dimers formed by Mg/Ca and one of these ligands were constructed for EDA calculations. All of the structures dimers were geometry-optimized by using MP2/cc-pVTZ. The optimized structures were used as a starting point for rigid distance scan, where the distance between the monomers was varied by moving the ion along the shortest vector between the monomers. In addition, a common binding mode for Mg/Ca and carboxylate group is monodentate binding, which is not the lowest energy structure in gas phase. The monodentate structure was constructed by fixing the angle C=O…M at 150 degrees, where M stands for Mg/Ca ion. The optimized dimer structures are shown in Figure 9**.** The interaction energies for the entire distance scan were calculated by using MP2/CBS level of theory, extrapolated from MP2/aug-cc-pV[Q5]Z. Energy decomposition was calculated by using SAPT2+/aug-cc-pVTZ, ALMO/HF/aug-cc-pVDZ, ALMO/ωB97X-D/def2-TZVPPD. In addition, the interaction energy for the optimized structures was also calculated by using

CCSD(T)/CBS, $E_{\mathrm{CCSD(T)/CBS}} = E_{\mathrm{MP2/CBS}} + E_{\mathrm{CCSD(T)/aTZ}} - E_{\mathrm{MP2/aTZ}}$. Gaussian 09 was used for geometry optimization. The Psi4 package was used for interaction energies and SAPT EDA, while Q-Chem 4.2 was used for ALMO-EDA.

The hydration free energies were calculated by FEP with Bennet acceptance ratio (BAR). The simulation box of (24.6 Å)^3 contained one metal ion and about 500 water molecules, and the density was set to 1.0 g/cm^3. For the alchemical transition, the coupling between the ion and its environment is parameterized by λ, where $\lambda = 0$ means completely decoupled and $\lambda = 1$ means full interaction. Two λ parameters were used for vdW and electrostatics. The vdW interaction was gradually turned on before the electrostatics interaction was turned on. Softcore vdW function was used to improve the convergences for vdW is nearly decoupled. In total, 20 thermodynamic states were simulated in the NVT ensemble at 300 K. The coordinates were saved every 5 ps for the BAR calculation. Details of other simulation settings are described in a previous papre.[58] All the MD simulations were performed by using TINKER-OpenMM.[11,102]

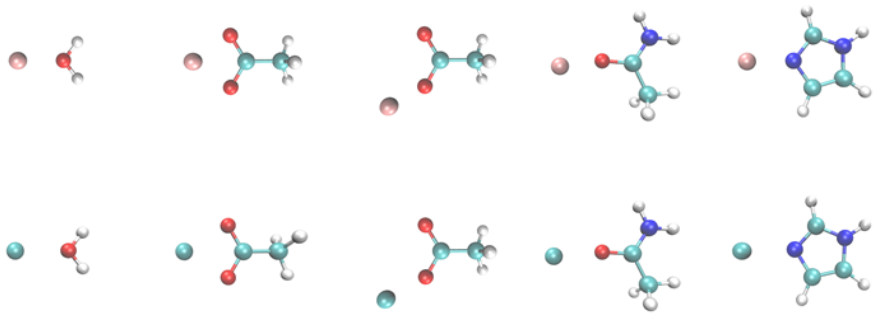


Figure 9: Optimized structures of the metal ion-ligand dimers.

The isolated pink and green spheres are Mg2+ and Ca ions. White, red, green and blue spheres in the ligand represent H, O, C and N atoms, respectively.

2.5 Refinement of Mg/Ca parameters

Model	Dimer interaction energy (kcal/mol)				
	AMOEBA	CCSD(T)	MP2	SAPT2+	ALMO
Mg-Water	-83.40	-83.12	-83.07	-80.79	-80.28
Ca-Water	-56.12	-57.15	-57.56	-55.64	-53.96
Mg-Acetate	-392.59	-387.79	-385.99	-390.61	-382.31
Ca-Acetate	-330.15	-329.84	-329.68	-335.40	-320.89
Mg-Acetamide	-141.36	-144.84	-143.50	-140.34	-143.58
Ca-Acetamide	-100.29	-108.94	-108.65	-105.63	-105.42
Mg-Imidazole	-135.97	-139.04	-138.28	-138.55	-133.71
Ca-Imidazole	-99.49	-97.26	-97.63	-99.21	-90.79

Table 6: Equilibrium dimer interaction energies calculated by AMOEBA and different QM methods.

The previous AMOEBA Mg/Ca force fields were developed using a relatively low level of QM method from current standard (MP2/aTZ). This small basis set was shown to underestimate the Mg/Ca-water binding energies about 5 kcal/mol. Therefore, the previous AMOEBA force field also underestimated the interaction energies. The hydration free energies calculated from the previous work have large uncertainties. Here, the Mg/Ca parameters were optimized by using high level MP2/CBS data. The QM and AMOEBA interactions energies were compared in the top panels of Figure 10. Good agreement was found for the entire interaction energy curve. The errors in the interaction energies at equilibrium distance are ~1 kcal/mol. The errors in the equilibrium distance

are very small for Ca and slightly larger for Mg. The entire interaction energy curve is accurately predicted for Ca. The largest error for Ca is 3.1 kcal/mol at the largest separation distance. The error for Mg is –3.9 kcal/mol at the distance of 1.55 Å. The new parameters can better reproduce the experimental hydration free energies.[73]

3. ENERGY DECOMPOSITION ANALYSIS FOR DIMERS OF MG/CA AND SMALL MOLECULES.

3.1 Water

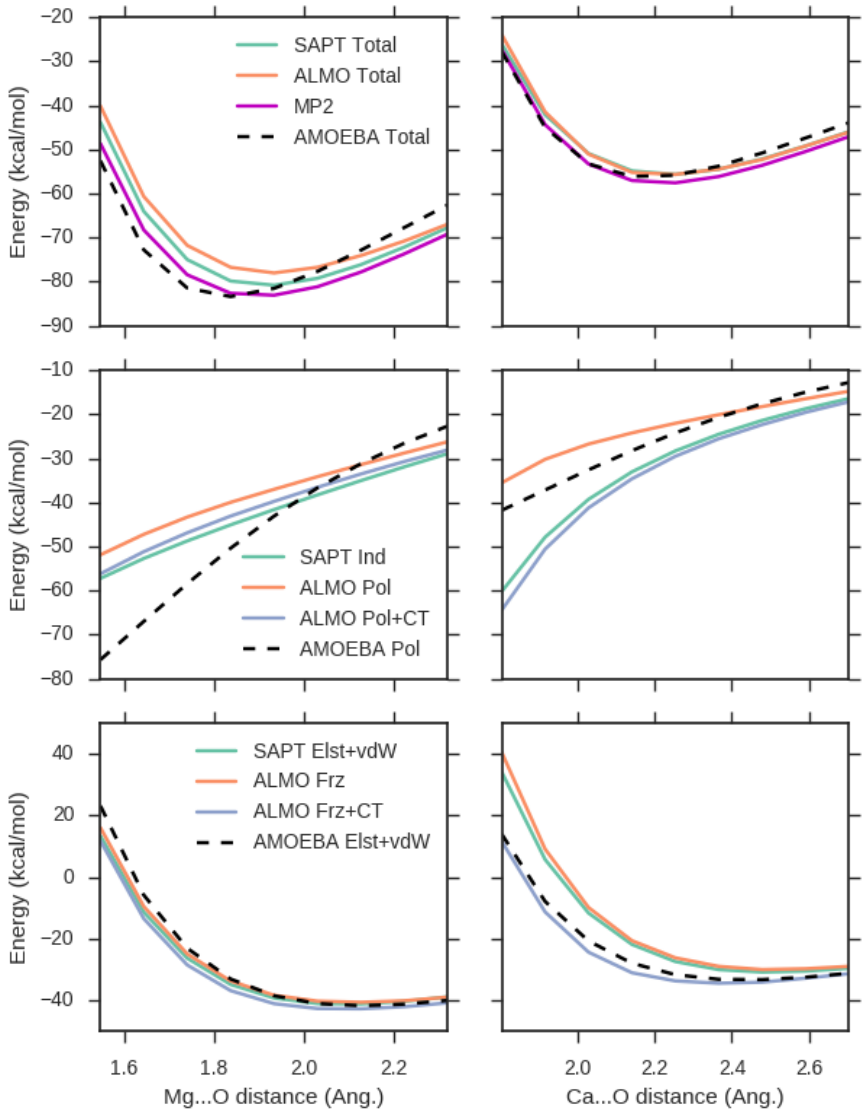


Figure 10. QM and AMOEBA interactions energies for Mg-water and Ca-water complexes.

"Ind", "Pol", "Elst", "Frz" and "CT" stand for induction, polarization, electrostatics, frozen core, and charge transfer, respectively. By convention, induction includes polarization and charge transfer; frozen core includes electrostatics and vdW.

First, ion-water interactions are investigated. Water is important for ion binding because it serves as a reference state and many binding sites contain coordinating water molecules.

The accuracy of ALMO and SAPT was evaluated based on the total interaction energies. ALMO and SAPT are very close for the Ca...water dimer, while some difference exists for Mg…water dimer. In both cases, ALMO and SAPT are less attractive than the reference MP2 value. The errors of ALMO and SAPT for Ca…water are about 2 kcal/mol; for Mg…water dimer, SAPT has an error of 2~5 kcal/mol in repulsive region, while the corresponding error for ALMO is 5~9 kcal/mol. The Mg…water dimer has a similar trend: ALMO and SAPT have smaller interaction energies than the reference value, and the error is even larger than in the case of Ca. The errors are due to both the basis set and the Hamiltonian. In this work, the ωB97X-D functional was used for ALMO. More accurate functionals that belong to the same family have been reported, such as ωB97X-V and ωB97M-V.[103]

The equilibrium interaction energy calculated by AMOEBA is close to the reference MP2 values, while the equilibrium distance is slightly shorter than that from QM. In addition, AMOEBA is less accurate for Mg than for Ca, which is also the trend observed for ALMO and SAPT.

To understand the errors of AMOEBA, energy components from AMOEBA were compared to those from SAPT and ALMO (Figure 10). For short range interactions are dominated by Frz, while interactions at longer distances consist of comparable contributions from Frz and Ind. ALMO total induction energy (Pol+CT) is close to SAPT induction energy, while ALMO Frz interaction is close to SAPT electrostatics plus vdW. This is consistent with the definition of induction energy in SAPT. Therefore, reliable energy decomposition can be obtained for the ion-water dimers.

There is a significant charge transfer energy for the Ca…water interaction. At equilibrium distances, Ca…water dimer has larger CT energy (-7.5 kcal/mol) than Mg…water dimer (-2.7 kcal/mol). A similar trend was obtained by using a second generation ALMO analysis.[104] However, this trend contradicts the chemical intuition that Mg should have a larger CT energy because the amount of charge transfer measured by electrons is larger for Mg…water than for Ca…water does. A previous study found that ALMO/ωB97X-D3 ranks alkali metal ions as Li, K, Na in order of decreasing ion-water charger transfer energy, the sequence according to chemical intuition should be Li, Na, K.[105] Therefore, the CT energy for Mg…water and Ca…water computed by ALMO may not be physically meaningful.

The AMOEBA polarization energy is compared to ALMO induction energy and polarization energy. AMOEBA polarization is closer to the ALMO polarization energy than the total induction energy. Accordingly, AMOEBA Frz is closer to ALMO Frz+CT than ALMO Frz energy. Therefore, the CT energy is implicitly absorbed into Frz energy in AMOEBA. The AMOEBA short-range polarization is more attractive than its QM counterpart. Therefore, the short-range damping function in AMOEBA may need to be improved. Generally, AMOEBA performs better for Ca than for Mg in terms of the accuracy for polarization energy. However, because of the uncertainty in CT energy from EDA, the meaning of this agreement is unclear.

Frz energy can be further separated into electrostatics and vdW. The results (not shown in Figure 10) indicate that AMOEBA has accurate vdW interaction for the Mg...water dimer. The AMOEBA electrostatics interaction for Mg…water is accurate expect for short distance, with a largest error -11 kcal/mol. For Ca…water, the largest error in the AMOEBA electrostatics interaction is -31 kcal/mol. The error in

electrostatics is much larger than the error in Frz, which indicates error cancelation between the buffered vdW interaction and electrostatics interaction.

3.2 Acetate

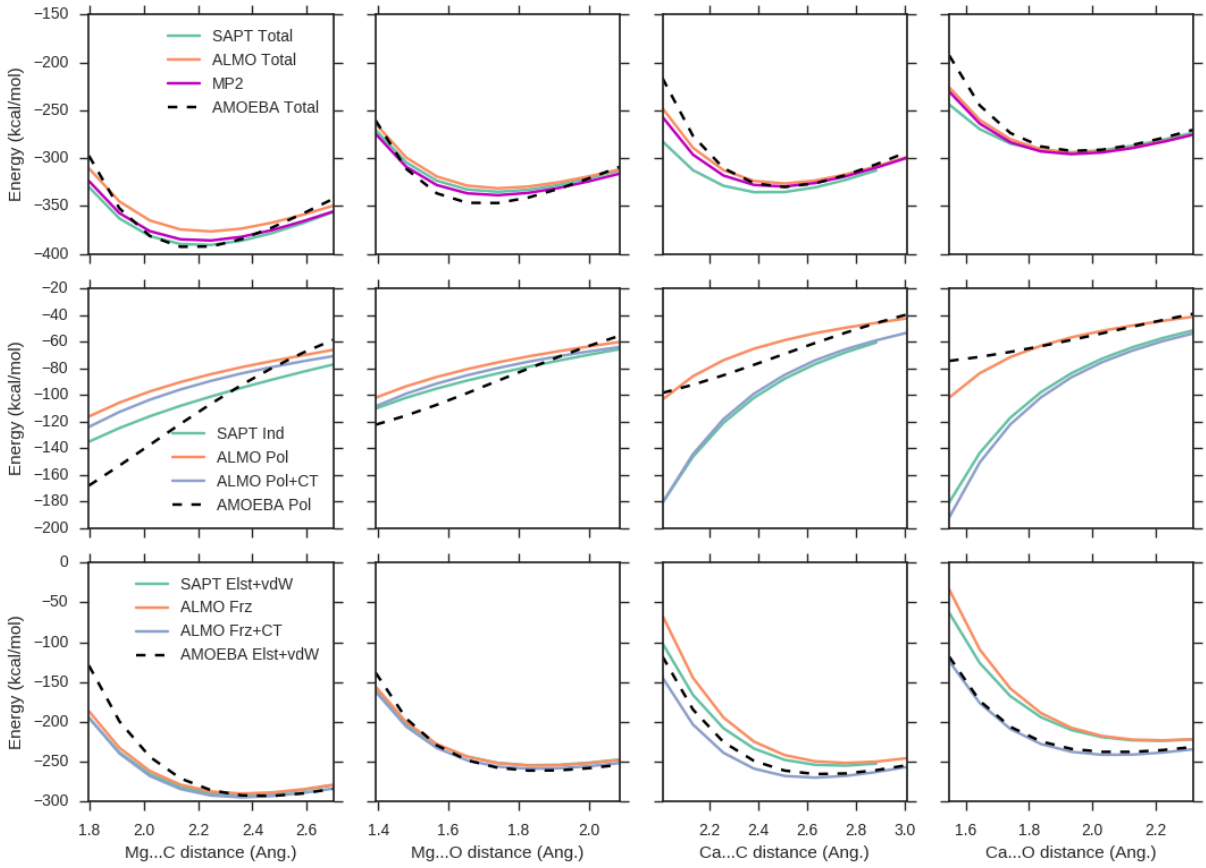


Figure 11: QM and AMOEBA interactions energies for Mg-acetate and Ca-acetate complexes.

See Figure 10 caption for explanation of the energy components.

Carboxylate groups are also very common in Mg/Ca binding pockets. In Ca-binding proteins, there are usually 3-4 carboxylate groups. The interaction between

Mg/Ca and the carboxylate groups are also more challenging for classical force fields compared to the interactions with water, since carboxylate groups are charged and interact more strongly.

There are two binding modes for Mg/Ca…acetate dimers (Figure 11). AMOEBA accurately reproduces the equilibrium distances and interaction energies for all four dimers. The errors in equilibrium interaction energies are -5.2/-0.6 kcal/mol for the bidentate binding mode of Mg/Ca, and -8.1/+3.3 kcal/mol the monodentate binding mode of Mg/Ca. The relative errors are very small considering the magnitude of the total interaction energy. The errors for the relative interaction energy between Mg and Ca are -2.9/+3.9 kcal/mol for the bidentate and monodentate binding mode. AMOEBA interaction energies at short distances are smaller than the QM interaction energies. The accuracy for Ca is better for than for Mg, which is consistent with the results for the ion-water dimers. In addition, AMOEBA performs better for the bidentate binding mode than for the monodentate binding mode.

Due to the strong electrostatic interactions between carboxylate and the metal ion, polarization component is significant. The AMOEBA polarization energy curves have different slope compared to the ALMO polarization energies. For Mg…Ac-, the AMOEBA curve is steeper. The Mg…Ac polarization energy in AMOEBA is smaller than ALMO for longer distances and larger than ALMO for shorter distances. Also, AMOEBA suffers from a larger error for Mg…Ac- bidentate binding than Mg…Ac- monodentate binding. The larger error in polarization for Mg…Ac- bidentate binding is partly compensated by the stiffer vdW interaction. Therefore, the vdW interaction AMOEBA helps correct the over-polarization error.[73]

For Ca…Ac-, the agreement between AMOEBA and ALMO polarization is much better, although the curvature across the range of interaction distances is not perfect. For

Ca…Ac- bidentate binding, AMOEBA polarization is smaller than ALMO around the equilibrium distance and larger than ALMO at longer distance. The curvatures for polarization in the monodentate and bidentate binding modes are similar to each other, although the range of distances was calculated on different pairs of atoms and was chosen to be near the respective equilibrium distances for each binding mode. For Ca…Ac- monodentate binding, AMOEBA polarization is accurate around the equilibrium distance, and larger than ALMO at longer distance. The conformations of the two binding modes could interconvert between each other during the simulations, and the relative energy between the two binding modes are important for modeling the coordination geometry. The fact that AMOEBA is more accurate for one binding mode than the other indicates that the carboxylate interactions are challenging for classical force fields.

The Frz energy in AMOEBA agrees reasonably with the ALMO Frz+CT energy for all ion-acetate dimers. For Mg, AMOEBA electrostatics energy matches the SAPT electrostatics component. For Ca, AMOEBA electrostatics is less attractive than SAPT at short distances. The error in electrostatics is compensated by the vdW interactions. The AMOEBA vdW interaction for Mg…Ac- bidentate binding mode is also more repulsive than vdW+CT energy from EDA. The above comparisons show that the vdW component in AMOEBA not only captures the short-range electrostatics and charge-penetration interactions, but also helps reproduce the relative binding energy between bi- and mono-dentate binding.

3.3 Acetamide

The second most common ligands in Ca-binding proteins are the carbonyl groups from either the protein backbone or Asn/Gln side chains. Acetamide is used as a reduced

model for the carbonyl groups. The comparison of total interaction energies and energy components are shown in Figure 12. Overall, the performance of AMOEBA for ion-acetamide is worse than the performance for ion-water and ion-acetate dimers. AMOEBA underestimated the equilibrium distances for Mg…acetamide and Ca…acetamide by 0.1~0.2 Å. The error in equilibrium interaction energies are 1.3 and 8.4 kcal/mol for Mg…acetamide and Ca…acetamide, respectively. This problem is caused by the overall weaker polarization and electrostatics energies in AMOEBA. Optimization of vdW parameters were unable to compensate for the error at the entire range of distances. Therefore, sacrifice in equilibrium distance and/or interaction energies was necessary.

Ca…acetamide dimer have energy components similar to the Ca…water dimer. The CT contribution amounts to 13-15 % of the total interaction energy, and Frz and Pol contributions account for the rest. To absorb the CT and charge penetration effect into vdW interactions, the AMOEBA vdW interaction in Ca…acetamide in should be twice as much as in Ca…water. This cannot be achieved since the vdW parameters for the water oxygen and the carbonyl oxygen should be consistent. Apart from the poor description of CT, the polarization energy in AMOEBA is also too repulsive even at short range.

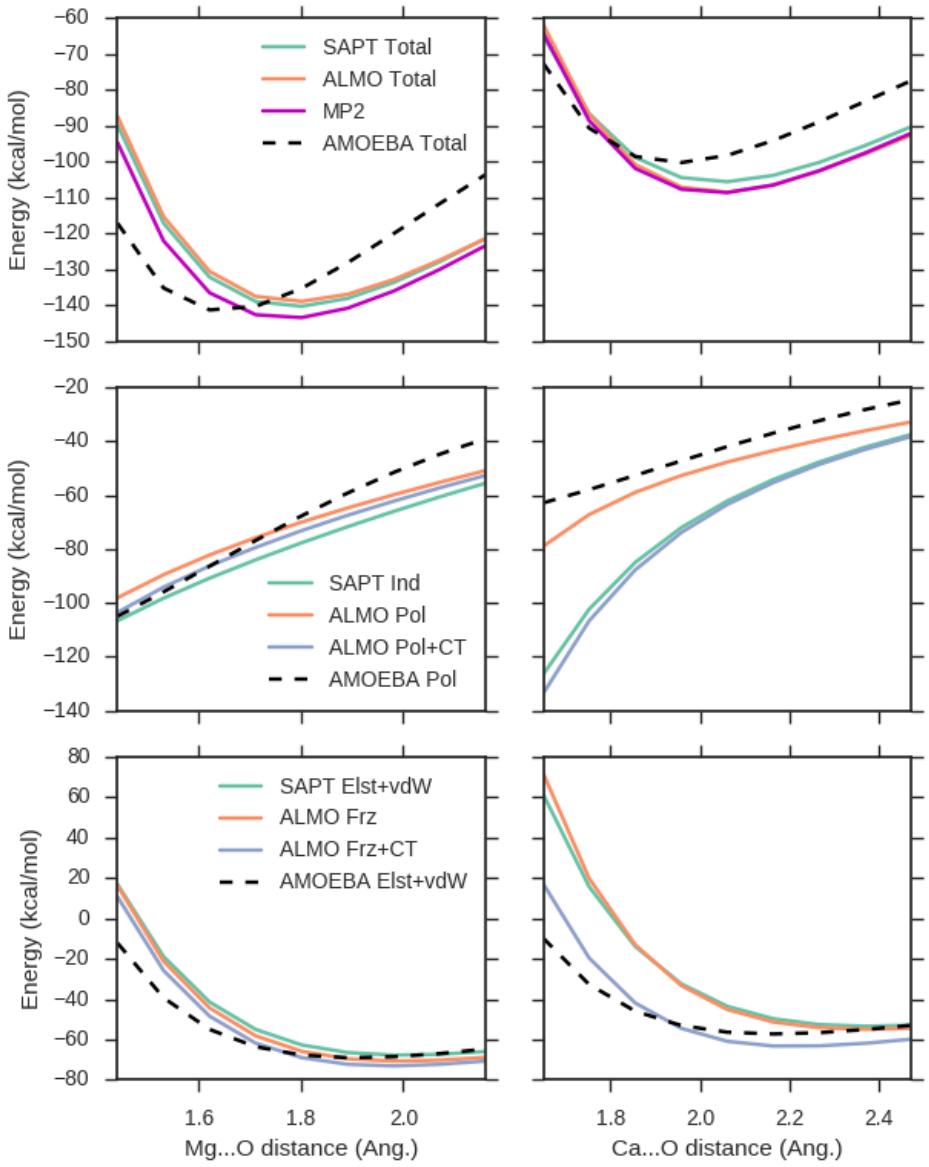


Figure 12: QM and AMOEBA interactions energies for Mg-acetamide and Ca-acetamide complexes.

See Figure 10 caption for explanation of the energy components.

3.4 Imidazole

Imidazole is also an import residue for metal ion binding. The AMOEBA energies for Mg/Ca…imidazole have a similar trend to those for Mg/Ca…acetamide. The interaction is too soft at short separation and too strong at long separation (Figure 13). The error in equilibrium distances is 0.1~0.2 Å and the error in the total interaction energy is around 2%. The Frz energies are reasonably reproduced, and the polarization energies are more problematic. Mg…imidazole has a larger error than Ca…imidazole, which can be attributed to polarization. The Ca…imidazole polarization energy agrees better at short separation, which is because Ca has a larger size than Mg.

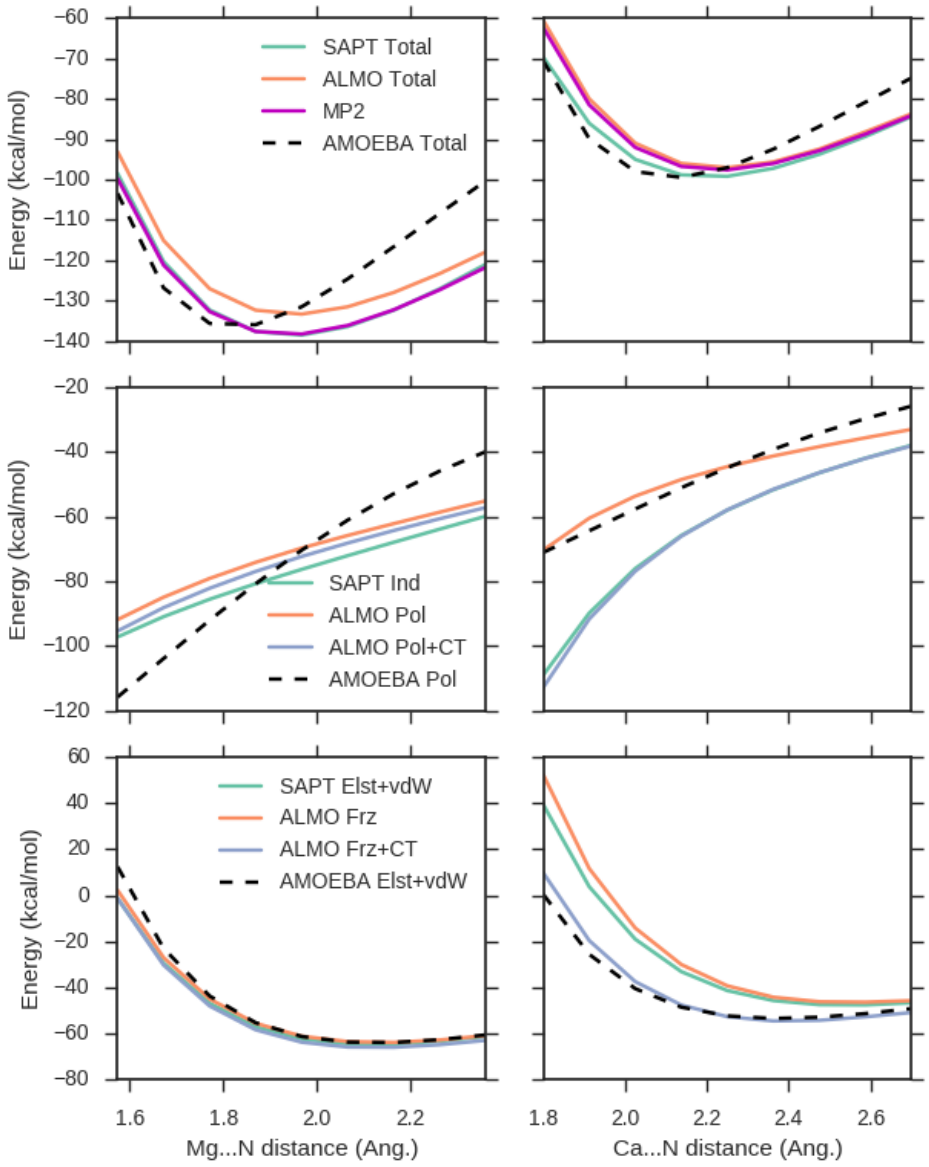


Figure 13: QM and AMOEBA interactions energies for Mg-imidazole and Ca-imidazole complexes.

See Figure 10 caption for explanation.

4. MANY-BODY EFFECT IN ION-PROTEIN BINDING.

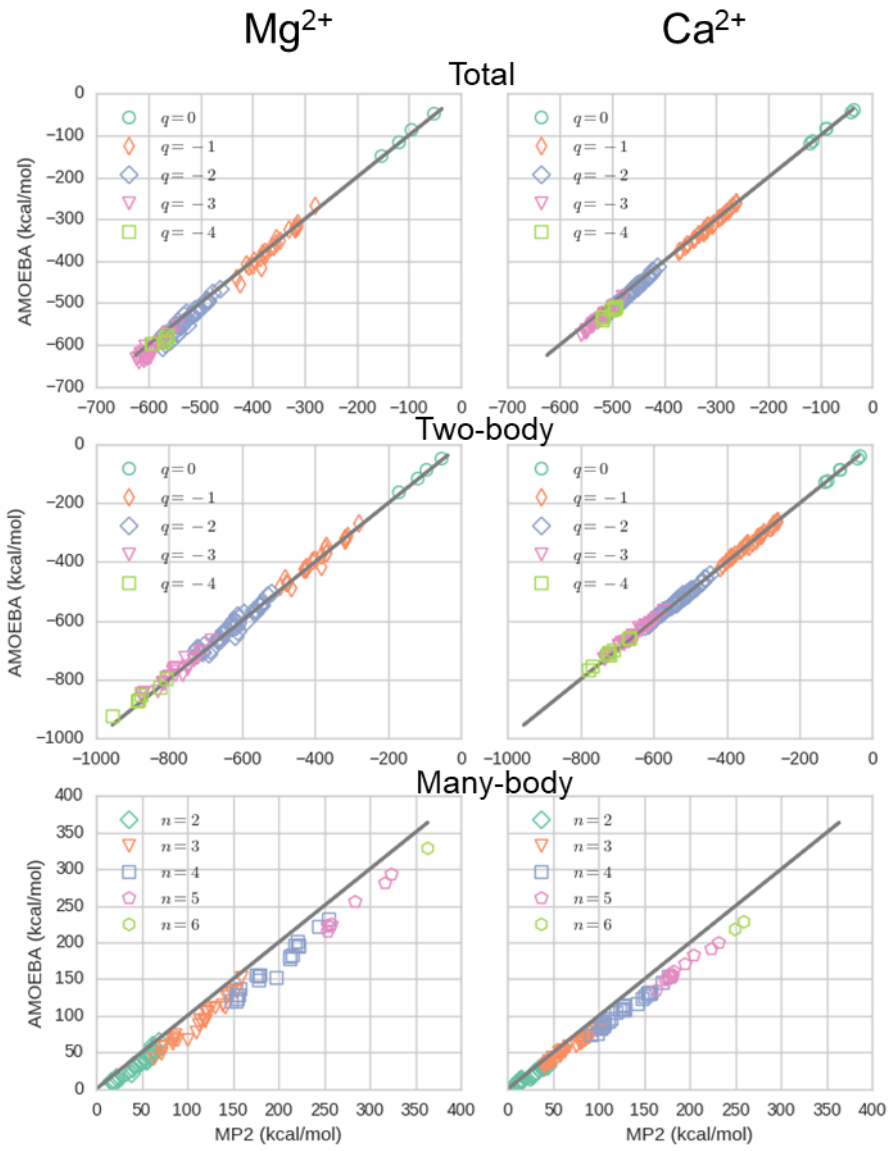


Figure 14: MP2 and AMOEBA interactions energies for model compounds of calcium-binding pockets.

Each model compound contains one Mg/Ca and one or more molecular fragments representing the coordination in the binding pockets. "*q*" and "*n*" denotes the net charge (excluding the charge of the metal ion) and the number of ligands in the model compounds, respectively. The figure was adapted from referennce.[73]

The metal ion binding pocket generally consist of multiple carboxylate groups. Because of the strong charge-charge interactions, there should be significant many-body effect. It is known experimentally that the binding pocket may have a different selectivity from each of the ligand, so the many-body effect may also be important for ion selectivity. The many body effect was analyzed for a set of model compounds for the binding pockets (Figure 14). AMOEBA total energies agree very well with the reference values. For larger compounds, AMEOBA interaction is overestimated. Relatively speaking, better accuracy was achieved for Ca compounds, which is consistent with the observation for the dimer interaction energies. The total interaction energies are correlated with the net charge, although the relationship is not monotonic. The pockets with -3 net charge have lower energy than the pockets with -4 and -2 net charges, which means that the binding pockets with four carboxylate groups are less favorable if there is no protein matrix.

It seems very intuitive that due to the charge-charge repulsion, the binding pockets with -4 net charges are less stable than those with -3 net charges. However, the role of many-body effect on the stabilization has not been recognized. The two-body energy in Figure 14 suggest that without the many-body effect, the pocket with -4 net charges will be the most stable. The many-body effect destabilizes the pocket with -4 net charges because the contribution of charge transfer and polarization to many-body effect are both positive. When only two-body interactions are considered, the Mg compounds is 178 kcal/mol lower in energy than Ca compounds, which makes the binding pocket to be Mg-selective. After considering the many-body effect, the Mg compounds is only 73 kcal/mol lower in energy than Ca compounds, which falls within the range of the difference in hydration free energies and makes it possible to have different selectivities.

The AMOEBA and QM many-body energies are compared in Figure 14. As expected, Mg-pocket has larger many-body energy than Ca-pocket. Good correlation between AMOEBA and QM was observed. However, the correlation for Mg is worse than for Ca, and AMOEBA systematically underestimated the many-body energies. The nearly linear relationship between AMOEBA and QM suggest that an improved polarization model may better predict the many-body energies. In addition, the small error in many-body energies may be caused by the missing charge transfer effect. Detailed analysis of the source of errors reveals all the data points with errors greater than 30 kcal/mol are related to the interaction with acetamide.

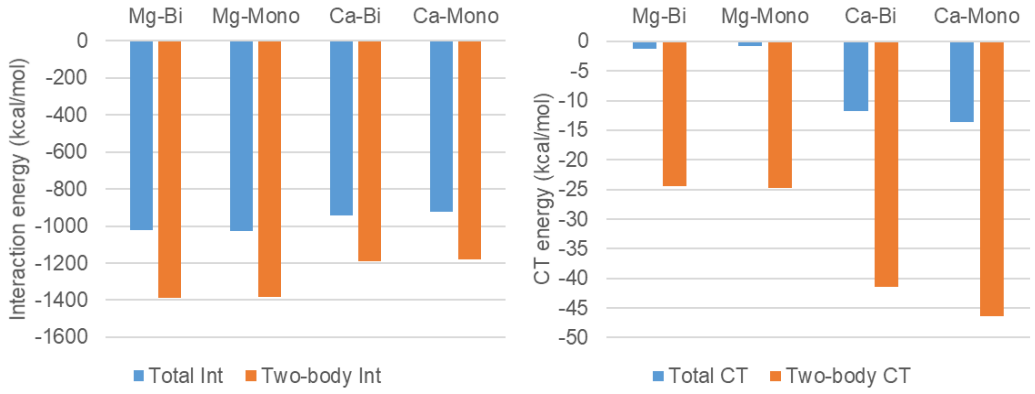


Figure 15: Interaction energy and charge transfer energy in ion binding.

"Bi" and "Mono" indicate bidentate binding mode (from short simulations of EF hand with Mg2+) and monodentate binding mode (from crystal structures of EF hand, PDB codes: 5cpv, 1b8l, 1b8c).

The effect charge-transfer was further quantified by EDA. The many body effect in both total energy and CT is shown Figure 15. In both cases, the interaction between ligands is much smaller than the ion-ligand interaction, and the ligand-ligand interaction is excluded to give a better picture of metal ion interactions. The many-body energies contributed to -35% and -27% of the total interaction energy for Mg and Ca, respectively.

The many-body effect is even stronger for CT. The total CT energy in Ca compounds is only ~30% of the two-body CT energy, which means a decrease of 70% due to the many-body effect. For Mg compounds, the total CT energy is only 3-5% of the two-body CT energies. This seems unphysical because the total CT energy in Mg compounds is smaller than in Mg-water dimer. More robust EDA method may help resolve this issue. The CT many-body effect in Ca compounds are ~30 kcal/mol, which account for ~10% of the total many-body energies. According to chemical trends, the ratio of CT many-body energy should be similar or larger for Mg, which amounts to ~40 kcal/mol. The magnitude of the CT many-body energy is very close to the error in AMOEBA many-body energy. This suggest that CT term may be a possible way to improve the description of the many-body effect..

5. SUMMARY

The AMOEBA Mg and Ca parameters were optimized by using high-level QM data on ion water dimer interaction energies. The new parameters better reproduced both the total interaction energy and each component and the solvation free energies (within 0.5 kcal/mol of experimental values).

Energy decomposition analyses of Mg and Ca interactions showed that AMOEBA is accurate for interactions with water, acerate or Ca. Some weaknesses of AMOEBA were also identified for short-range polarization and charge transfer interactions. Since Mg has stronger charge transfer than Ca, AMOEBA is less accurate for Mg than for Ca.

The many-body effect revealed by many-body expansion demonstrated that the many-body effect is important for the Mg/Ca selectivity in protein-ion binding. Mg compounds have much larger many-body energy than Ca compounds. This explains the

lower Mg binding affinity in many proteins. Therefore, the many-body effect should be included in force fields, either implicitly or explicitly, for simulations of Mg and Ca binding.

Chapter 4: Mechanism of ion selectivity in calcium-binding proteins

Note: this Chapter is based on a published paper[106] for which I conducted the experiment, analyzed the data and wrote the paper, Jing, Z., Liu, C., Qi, R. & Ren, P. Many-body effect determines the selectivity for Ca^{2+} and Mg^{2+} in proteins. Proc. Natl. Acad. Sci. U. S. A. 115, E7495-E7501, doi:10.1073/pnas.1805049115 (2018).

1. INTRODUCTION

Ca2+ is a second messenger for the regulation of many biological cellular processes throughout the cell life cycle. Ca2+ couples with calcium-binding receptors on the outside of cell membrane to regulate the transportation of calcium into a cell and to active hydrolytic enzymes and cell adhesion molecules.[77,107] Mg2+ is critical in metabolism since ATP must be bound to Mg2+ to have biological activity.[108]

To achieve their functions, calcium-binding proteins selectively bind Ca^{2+} against a background of competing ions such as Mg^{2+}. The mechanism for Mg^{2+}/Ca^{2+} selectivity is unclear. Mg^{2+} is a better charge acceptor and therefore is expected to interact favorably with highly charged binding pockets in calcium-binding proteins. However, these proteins bind Ca^{2+} with high selectivity. Explanations for the EF-hand selectivity include (1) the snug-fit mechanism, [109,110] i.e. the bi-pyramidal geometry of the binding pocket can perfectly fit Ca2+, but not Mg2+; (2) solvation penalty, i.e. Mg^{2+} has large solvation energy which makes the binding of Mg to proteins unfavorable. [111,112] The geometric argument cannot explain EF-hand mutants with octagonal geometry also favor Ca^{2+} ion.[113] The solvation penalty argument, although correct, does not offer any mechanistic insight but rather a phenomenological explanation. It does not answer the question why the interaction of Mg^{2+} with carboxylate groups is weaker than the solvation. Besides, the

solvation penalty argument does not explain the differential selectivity in different proteins. Many EF-hand proteins have optimal binding affinity for Ca^{2+} and weaker binding affinity for Mg^{2+}, Sr^{2+} and Ba^{2+}.[114] The CheY protein, which is similar to the EF hands in the chemical composition, has similar binding affinities for the series of divalent ions.[115]

In fact, there are many proteins that favor Ca2+,[107,111,116-119] including a Mg-binding protein, Ribonuclease H1.[120] By comparison, Mg-specific proteins are uncommon. There have been many studies on the specific factors contributing to calcium-selectivity. The protein environment was shown to stabilize the binding site geometry.[111,121-123] The geometric constraint or deformation[124-126] is also frequently discussed, while the number of negatively charged residues can affect the binding affinities through electrostatic repulsion.[127-129] The roles of polarization[130-135] and quantum effects[126,136,137] have also been pointed out. These studies provided clues on metal ion selectivity in some proteins, but does not explain the general trend of Ca/Mg selectivity in a wide range of proteins and inorganic ligands.

In this work, quantum mechanical calculations and polarizable force field simulations were used to decipher the mechanism for calcium-selectivity.

2. QUANTUM CHEMICAL STUDY OF PROTEIN-ION INTERACTIONS

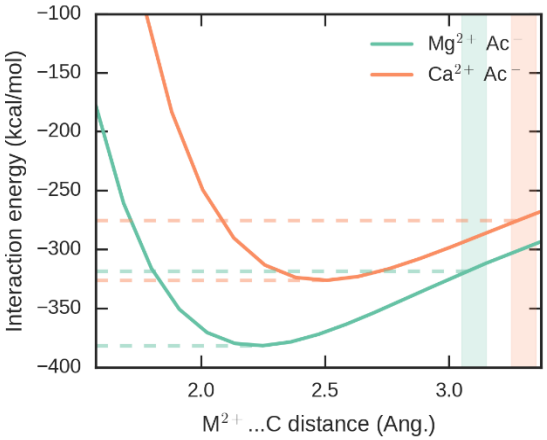


Figure 16: Interaction energy between Mg2+/Ca2+ and acetate as a function of the separation distance.

The interaction energies are calculated by RI-MP2/def2-QZVPPD. The shaded areas in green/orange denote typical distances between Mg2+/Ca2+ and carboxylate groups in PDB, respectively.

Mg2+ has much stronger interaction with acetate than Ca2+ does in vacuum (**Figure 16**). The difference between the interaction energies for the Mg-acetate and Ca-acetate dimers is as large as 55 kcal/mol. In protein environment, due to repulsion between negative charged residues and restraint from the protein scaffold, the carboxylate groups cannot adopt the most favorable configuration as in the isolated dimer. When this geometric effect is considered, the interactions for both Mg and Ca are weakened and their difference reduces to ~40 kcal/mol, but this is still significant. The difference between Mg2+ and Ca2+ hydration free energies is 77.8 kcal/mol[138] and the difference between ion-water cluster interaction energies calculated by the same method (RI-MP2/def2-QZVPPD) is 72.3 kcal/mol. If the interactions between Mg/Ca and carboxylate groups are additive, an EF-hand binding pocket with four Asp/Glu residues would favorably bind to Mg by a large margin.

The stronger dimer interaction for Mg does not transfer into favorable binding affinity with simple ligands or proteins. The chemical hardness of Mg is very close to that of Ca, implying their similar binding affinities for common ligands.[139] The binding free energies of acetate for Mg and Ca are -1.7 and -1.6 kcal/mol, respectively.[140] Multivalent ligands such as ethylenediaminetetraacetate (EDTA) tend to have higher affinities for Ca2+.[141] There are also many proteins, both EF hand and non-EF hand, that favorably bind Ca2+, while only one protein, the integrin I domain,[142] was found in the literature that has higher affinity for Mg2+. The similar affinities of acetate for Mg2+ and Ca2+ can be explained by the dielectric constant of the environment or the solvation penalty, while the higher affinity of multivalent ligands and proteins should be related to cooperativity. [143]

The quantum mechanical interaction energies and many-body expansion were calculated for several model compounds for protein-ion complexes (**Table 7**). The model compounds contain acetate, acetamide and water as a proxy for the binding pocket. Despite the simplicity of the reduced models, the total binding energies are in surprisingly good agreement with experimental binding free energies. The ΔE_{bind} of all the three EF-hand systems (1IG5/4ICB, 2LVJ/K, 1B8L/C) are smaller than or close to $\Delta\Delta G_{solv}$, which means Ca-selective after accounting for hydration penalty. This result suggests that the first-shell interactions play a major role in determining the selectivity, while the dielectric response of the environment to the first shell and the entropy contributions are similar for Mg and Ca binding. The second shell can still be important in the stabilization of the binding pocket, which is not considered here since only relative interaction energies between Mg and Ca complexes are of interest. The total interaction energies are decomposed into two-body (2B) and many-body (MB) interactions (**Table 7**). The 2B interactions are always favorable for Mg binding. However, the large

difference in MB interactions favors Ca, which compensates for the unfavorable 2B energy.

PDB	Model compounds	ΔE_{2B}	ΔE_{MB}	ΔE_{int}	ΔE_{bind}	Expt. $\Delta\Delta G_{bind}$
1IG5/4ICB	Ac-B-Ac-B-Ac-W-W	232.6	-165.6	67.0	66.0	-6.2
2LVJ/K	Ac-Ac-Ac-B-Ac-W	144.3	-73.1	71.1	71.3	-2.4
1B8L/C	Ac-Ac-Ac-B-Ac-W	180.3	-103.4	77.0	77.8	-1.6

Table 7. QM interaction energy (RI-MP2/def2-QZVPPD) and experimental binding free energy (kcal/mol).

Reported values are the difference between Mg-complexes and Ca-complexes, $\Delta E = E(\mathrm{Ca}^{2+}) - E(\mathrm{Mg}^{2+})$. The solvation free energy difference between Mg^{2+} and Ca^{2+} is $\Delta\Delta G_{solv} = \Delta G_{solv}(\mathrm{Ca}^{2+}) - \Delta G_{solv}(\mathrm{Mg}^{2+}) = 77.8$ kcal/mol. Acetate, acetamide and water are denoted by Ac, B, and W, respectively.

The large positive many-body energies are not explicitly modeled in additive fixed-charge force field, while they are included in polarizable force field since a large portion of them come from polarization.[138,144] The many-body energies calculated by AMOEBA and MP2 are compared for several model compounds for of binding pockets, including both Ca-binding and Mg-biding proteins (**Figure 17**). The many-body energies are generally larger when there are more carboxylate groups, and the many body energies for Mg-containing complexes are considerably larger than those of Ca-containing systems. A very good correlation is observed, although AMOEBA systematically underestimated the many-body energies by ~30 kcal/mol.

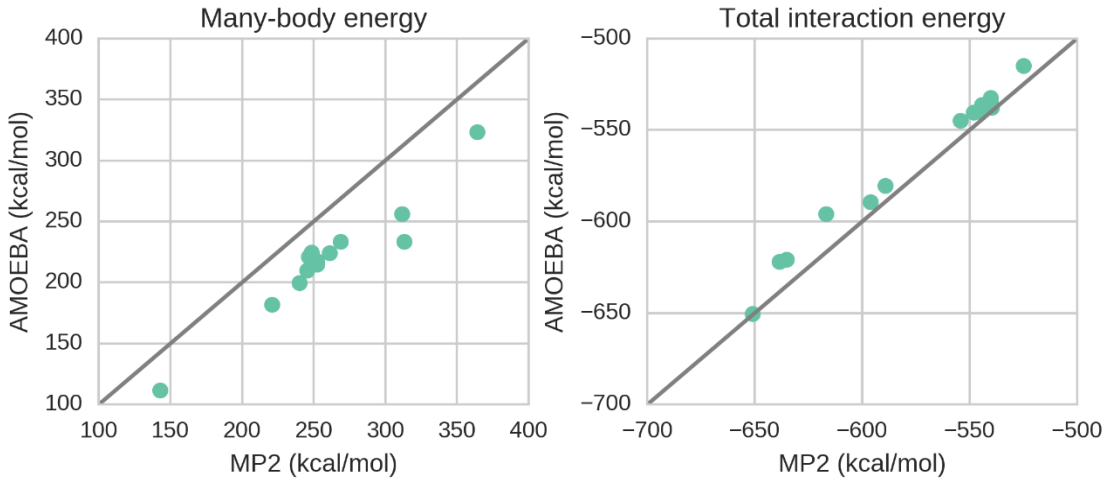


Figure 17. Many-body energies and total interaction energies in model compounds for Ca/Mg binding.

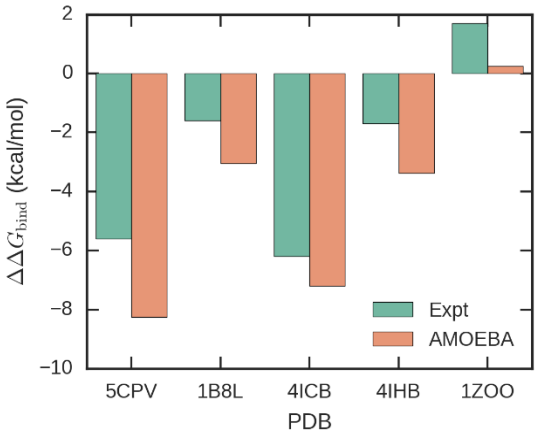


Figure 18. Experimental and calculated Mg/Ca relative binding free energies.

The standard deviations of all AMOEBA results are 0.2 kcal/mol.

PDB or ligand	Nr. of first-shell D/E		$\Delta\Delta G_{\text{bind}}(\text{Mg}\rightarrow\text{Ca})$	
		Expt.	AMOEBA	AMOEBA without polarization
5CPV	4	-5.6	-8.3	9.8
1B8L	4	-1.6	-3.1	10.4
4ICB	3	-6.2	-7.2	7.1
4IHB	2	-1.7	-3.4	4.6
1ZOO	1	1.7	-0.2	4.8
Acetate		-0.1	0.2	3.6

Table 8. Mg/Ca relative binding free energies for proteins and acetate (kcal/mol).

The standard deviations of all AMOEBA results are 0.2 kcal/mol.

3. THERMODYNAMIC DRIVING FORCE FOR ION SELECTIVITY

The importance of the many-body interactions was demonstrated by QM calculations. However, due to the relatively small system size and limited sampling, QM calculations cannot provide an accurate estimate of the thermodynamics of ion binding. MD simulations with AMOEBA was used to study the binding free energies since AMOEBA has reasonable performance for predicting the many-body energies. The calculated binding free energies of acetate for Mg and Ca are -2.5±0.2 and -2.7±0.2 kcal/mol, respectively. For comparison, the experimental binding free energies are -1.7 and -1.6 kcal/mol, respectively. Figure 18 shows the calculated and experimental relative binding free energy $\Delta\Delta G_{\text{bind}}(\text{Mg}\rightarrow\text{Ca})$ for five proteins, including two EF hand proteins (5CPV and 1B8L)[113] and calbindin D9k (4ICB),[117] the C2 domain of dysferlin (4IHB)[116] and integrin (1ZOO).[142] The $\Delta\Delta G_{\text{bind}}(\text{Mg}\rightarrow\text{Ca})$ calculated by AMOEBA are systematically lower by ~2 kcal/mol than the experimental values. Our previous work also shows that the Mg-ligand interactions in AMOEBA are weaker than QM.[138]

Nevertheless, the trend across different proteins is reproduced. There is a weak correlation between the number of first-shell Asp/Glu side chains, i.e. binding pockets with more Asp/Glu have higher selectivity for Ca. An exception is 1B8L/C, a Glu-Asp mutant of 5CPV. An explanation for this trend is that more negatively charged groups lead to more positive many-body energies for Mg, which overcomes the favorable dimer interaction. In the case of 1B8L/C, the enlarged binding pocket reduces the many-body energies, and thus shifts the equilibrium towards Mg. To verify this assumption, we carried out free energy simulations with polarization turned off. The results are listed in **Table 8**. As expected, the relative binding free energies $\Delta\Delta G_{\mathrm{bind}}(\mathrm{Mg}\rightarrow\mathrm{Ca})$ for all the proteins become positive, i.e. Mg is always the favored species. Also, pockets with more Asp/Glu have higher selectivity for Mg, which is wrong but consistent with chemical intuition. Indeed, the many-body energies are crucial in shaping the selectivity for Ca over Mg. Interestingly, the difference between 5CPV and its mutant 1B8L/C vanishes when there is no polarization. So, the geometric effect is also related to polarization. The reason is that polarization has strong distance dependence than electrostatic interactions. For 1ZOO, the bridging water between Mg/Ca and Asp137 was quickly released in non-polarizable simulations, providing one more Asp in the first shell. As a result, the relative binding free energy of 1ZOO is close to that of 4IHB. This water molecule is strongly polarized by Mg/Ca and Asp, with a polarization energy of -16.8 kcal/mol for Mg and -9.6 kcal/mol for Ca. The polarization energy of the water coordinating Ca^{2+} in the other four proteins ranges from -4.5 to -6.2 kcal/mol. Therefore, without polarization, the bridging water cannot be stabilized.

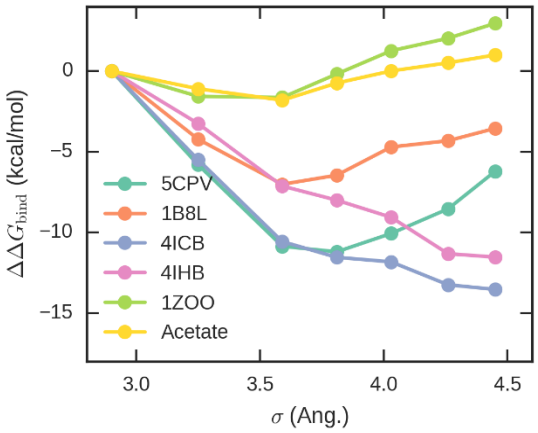


Figure 19: Relative binding free energy of artificial ions with different ion size.

The binding free energies are relative to that of Mg. The standard deviations of all results are 0.2 kcal/mol. The polarization parameter of Mg is used for all ions. 2.9, 3.59, 4.03, 4.45 Å roughly correspond to the vdW size of Ca2+, Sr2+, and Ba2+, respectively. These are not realistic parameters since the polarization parameters have not been optimized.

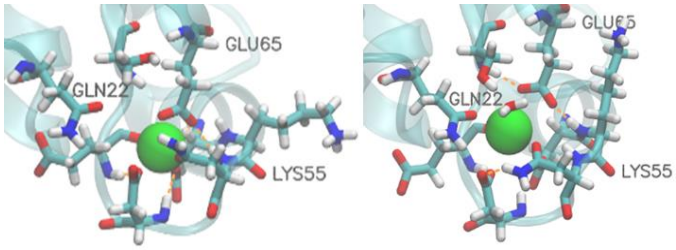


Figure 20: Structures of calbindin D9k (PDB code: 4ICB) bound to Mg2+ (σ = 2.90 Å) and an artificial ion with σ = 4.45 Å.

The favorable many-body energies for Ca2+ mainly arise from its larger size rather than its weaker polarizing ability, since the polarizability for both Ca and Mg are very small and their difference is negligible. If two ions have the same vdW size, the one with stronger polarizing ability will bind stronger to the pocket. To further elucidate the origin of the ion selectivity, we investigated the size dependence by using artificial metal ion with varied radius but the same polarization parameter. Relative binding free energies

were calculated for this series of ions (**Figure 19**). The five proteins show different patterns of dependence on the ion size. As the vdW radius σ of the ion increases, the binding energies of 4ICB and 4IHB decrease monotonically, while the binding energies of 5CPV, 1B8L and 1ZOO reach minima around σ = 3.6 Å, which corresponds to the size of Ca2+. Unlike 5CPV and 1B8L, 1ZOO prefers Mg2+ over the largest ion. These trends in the binding free energy correlate with the stiffness of the binding pockets. Both the two EF hand proteins and the integrin binding pockets have extensive H-bond networks, which prevent them from adapting to larger ions than Ca2+. While 4ICB is also an EF-hand protein, the ion can coordinate with an additional carbonyl group from Gln22 as its radius increases (**Figure 20**), which makes the binding pocket less specific for ionic size. The binding pocket of 4IHB is very flexible with only one H-bond between Asp22 and Asn41. Since 4ICB and 4IHB binding pockets are flexible to accommodate larger ions, the binding free energy decreases with the size of the ion due to the more favorable many-body energy. In addition, 1ZOO has only one carboxylate group in the first shell, and the many-body energy is less significant. So 1ZOO has the highest binding free energy for larger ions among all the proteins. On a final note, neutral mutation at the EF hand gateway position was found to maintain the Mg/Ca selectivity, while selectivity over larger ions becomes weaker.[128] This agrees with our results where the sub-optimal H-bond networks in mutants have weaker restraint on the size of the binding pockets, so the binding affinities for larger ions are enhanced.

4. SUMMARY

This work showed that the Ca/Mg selectivity of several proteins is determined by the positive many-body interactions. When there are multiple ligands in a binding pocket, the many-body energies dictate the ion selectivity. The larger ion Ca2+ has weaker many-

body interactions and thus higher binding affinities than Mg2+. In addition, proteins can avoid binding large ions by forming hydrogen-bond network that restrict the size of the binding pockets. Polarizable force fields can capture most of the many-body energies in Mg/Ca complexes and accurately reproduce the trend of the relative binding energies across different types of binding pockets.

5. COMPUTATIONAL DETAILS

Mg-bound and Ca-bound proteins structures, including 1IG5/4ICB, 2LVJ/2LVK and 1B8L/C, were selected to generate the model compounds for QM calculations. The first-shell residues were extracted and represented by small molecules including acetate, ethanol and acetamide. In the PDB structures, 2LVK has one more first-shell Asp residue than 2LVJ. To ease comparison, Asp53 in 2LVK was removed. The positions of hydrogen atoms were optimized by using DFT method B3LYP/6-31G(d) with other atoms fixed. The many-body energy decomposition was calculated by using MP2/def2-QZVPPD. No basis-set superposition error correction was applied for this large basis set. Gaussian 09 software was used for the geometry optimization, the Psi4 package was used for RI-MP2 calculations, and Q-Chem 5.0 was used for the implicit solvation calculation.

TINKER-OpenMM[145] was used for free energy calculations using AMOEBA.[145] AMOEBA force field with modifications of the Asp/Glu torsion was used.[138,146,147] Hydrogen atoms were added using PDB2PQR before the proteins were placed in solvent boxes. Salts were added to give a NaCl concentration of 0.15 M. The systems were minimized, heated and the density was equilibrated at 300 K and 1 atm in a total of 2 ns simulations. Free energy perturbation contains 5 intermediate steps to mutate from Mg to Ca. Production NVT simulations were performed for 6-ns at 300 K. Other standard simulation parameters are described in previous papers.[145]

Chapter 5: Quasi-chemical theory of potassium and rubidium solvation

1. INTRODUCTION

Ion-water interaction is an important factor for ion selectivity and protein-ion binding since most of the experiments and biological activities are in an aqueous environment. In addition, experimental thermodynamics data for ion hydration serve as a benchmark for theoretical methods, including ab initio QM methods and force fields.[148]

Quasi-chemical theory (QCT) analyzes ion solvation in terms of short-range interaction between ion and first-shell solvent molecules and long-range interactions between the first shell and its solvent environment.[148] The quasi-chemical species of ion and its first shell is used to calculate the long-range interactions. QCT has been a major theoretical method for the study of ion solvation because of the usefulness of the decomposition and the low computational requirement. The short-range interaction can be calculated by high-level QM methods, while the long-range interaction can be treated by much cheaper continuum solvation methods with little sacrifice in accuracy.

The thermodynamic cycle for QCT is shown in Figure 21. The hydration free energy of an ion M is calculated as the sum of free energy changes in five steps: (a) moving n water molecules from solution to gas phase; (b) the formation of ion-water clusters from ion and water monomers; (c) restraining the structure of the gas-phase ion-water cluster to be only inner-shell structures; (d) hydration of the inner-shell structure of ion-water clusters; (e) removing the restraint of inner-shell structure in solvent environment. Step (c) can be determined by the free energies of different structures. Step (e) can be determined by the population of different structures in ab initio simulations of ion solvation. Step (a) and (d) are usually calculated by continuum solvation methods.

Step (b) is the short-range interaction. The calculation of hydration free energy is summarized in Eq..

$$\Delta G_{\mathrm{hyd}} = -n\left(RT\ln\left(\frac{\rho_{\mathrm{water}}}{\rho_{\mathrm{gas}}}\right) + \Delta G_{\mathrm{pcm}}(\mathrm{water})\right) + \Delta G_{\mathrm{clst}}(n) - RT\ln p_n(\mathrm{inner}) + \Delta G_{\mathrm{pcm}}(n) + RT\ln p_r(n) \quad (5.1)$$

Where $p_{\mathrm{n}}(\mathrm{inner})$ is the probability of the gas-phase cluster to have an inner-shell structure, $p_{\mathrm{r}}(n)$ is the probability that n solvent molecules are within a distance r of the ion, and r is the cutoff distance for defining inner-shell structure. The values of $p_{\mathrm{r}}(n)$ for different ions have been determined from previous ab initio simulations.[149-151]

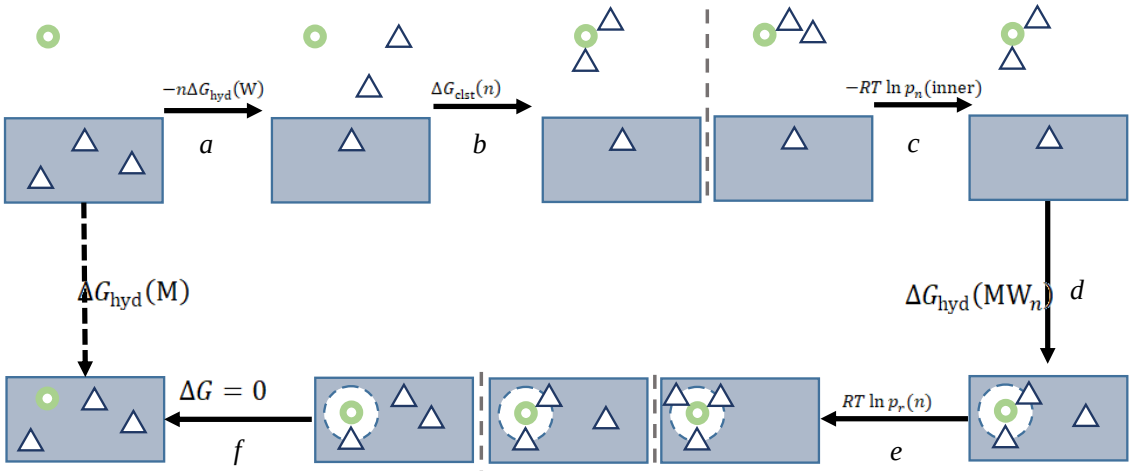


Figure 21: Thermodynamic cycle for ion solvation.

Green circle and navy triangle represent ion and water molecules, respectively, and the blue box represents bulk water. Dashed circle represents a sphere of radius r. Possible conformations of one thermodynamic state are separated by dashed lines.

In this work, various DFT methods are compared for their accuracy of ion interactions and their ability to predict hydration free energies based on QCT. An

emphasis is placed on K and Rb. Where K solvation has been well studied, there has been less investigations on the accuracy of QM methods for Rb. This work is the foundation for the study of K and Rb binding in potassium channels.

2. COMPARISON OF DFT METHODS

A dataset comprising ion-water dimers, ion-amino acids dimers and ion-organic molecule clusters was used to compare the gas-phase interaction energies.

The ion-water dimers include structures and 1.0x and 1.2x equilibrium distances. The ions include Li, Na, K, Rb, Cs, Cl, Br, I.

The ion-amino acids dimers include MP2/aug-cc-pVTZ optimized geometries. The amino acids are Ser, Gly and Gly2.

The Ion-organic ligand clusters are formed by one ion (K/Rb) and 1-3 same-kind or mixed ligands (methanol, formamide, N-methylamide). The geometries were optimized by MP2/aug-cc-pVTZ.

For the entire data set, the reference values were calculated by MP2/CBS + δCCSD(T).

The detailed results for ion-organic ligand dataset are shown in Figure 22. The summary of results for all data sets are shown in Figure 23 and Figure 24. For the convergence of basis set, aug-cc-pVDZ is found sufficiently close to the basis-set limit for almost all DFT methods. Among the DFT methods considered, B97M-V and MN15 have good accuracy for ion-water and ion-amino dimers, while SCAN-D3, PBE0-D3 and B3LYP-D3 are better for ion-organic clusters. In terms of computational efficiency, B97M-V (or ωB97M-V) is a few times slower than others when using RI method in Psi4; in AIMD simulations, hybrid DFTs are significantly slower for large systems because of

HF exchange. Therefore, B97M-V and SCAN-D3 are the only options for AIMD simulations.

Comparison between experimental and QM gas-phase interaction energies are shown in Figure 25. QM and experiment gas-phase binding energies show reasonable agreement. The difference between the best (MN15) and the "worst" DFT (B3LYP-D3) is smaller than experimental uncertainties. The differences between DFTs on K/Rb are much smaller than those on Li/Na.

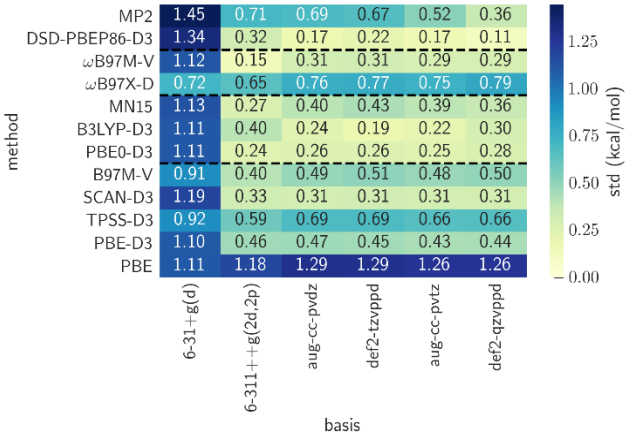


Figure 22: Error distribution of DFT methods and basis sets for the K/Rb-organic dataset. The reference value is calculated by MP2/CBS + δCCSD(T).

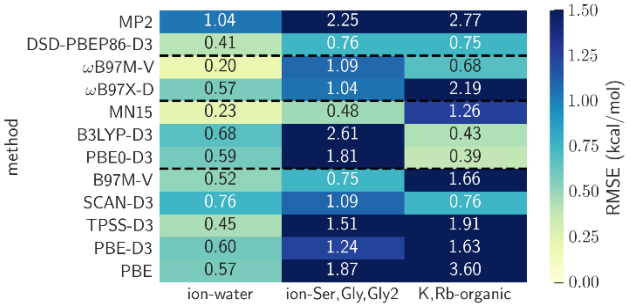


Figure 23: RMSE of DFT methods for different datasets. The reference value is calculated by MP2/CBS + δCCSD(T), and aug-cc-pvdz is used for DFT methods.

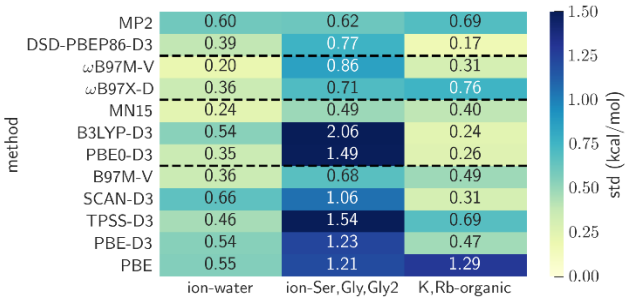


Figure 24: Error distribution of DFT for different datasets.

The reference value is calculated by MP2/CBS + δCCSD(T), and aug-cc-pvdz is used for DFT methods.

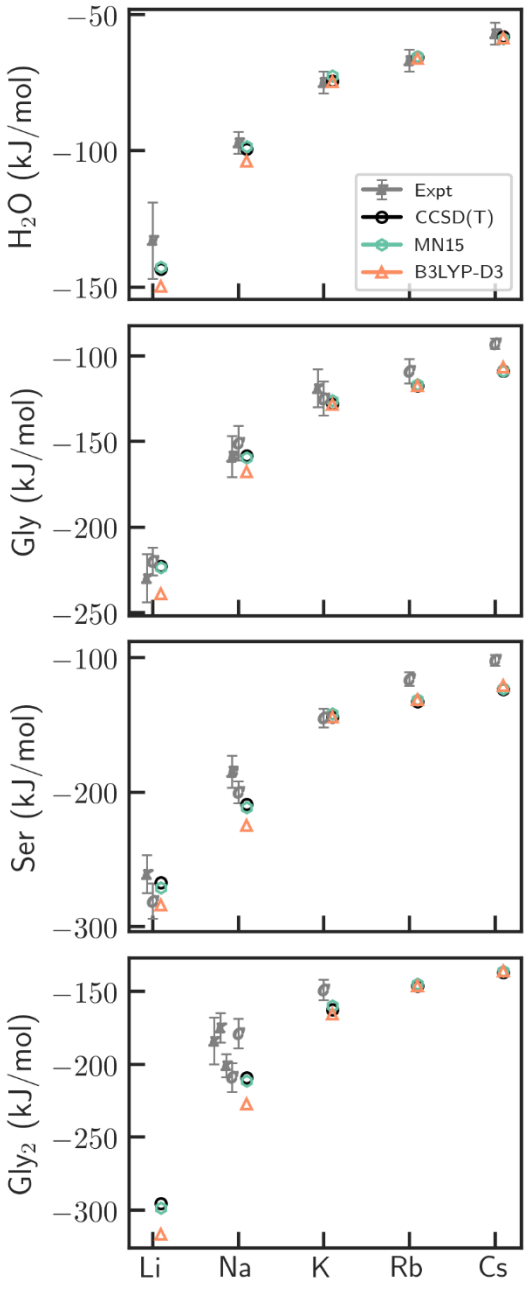


Figure 25. Comparison of DFT, CCSD(T) and experiments on gas-phase binding energy at 0 K.

C: threshold collision-induced dissociation; E: single temperature equilibrium; K: kinetic method results. The geometries and deformation energies are obtained by using

MP2/aug-cc-pvtz. CCSD(T) interaction energies are calculated by MP2/CBS + δCCSD(T)/aug-cc-pvdz, and aug-cc-pvdz is used for DFT.

3. ION HYDRATION FREE ENERGY

The short-range interaction $\Delta G_{clst}(n)$ in Eq. 1 is the most important term in QCT and it is sensitive to the QM method and the approach for free energy calculation. Given a QM or MM method, the free energy can be calculated through either harmonic approximation or MD free energy calculations. The computational details are described in the appendix.

The calculated hydration free energies using either harmonic approximation or MD for Rb+, K+ and F- are shown in Figure 26, Figure 27 and Figure 28, respectively. The calculated cluster interaction energies are compared with gas-phase experimental measurements and the final hydration free energies are compared to the results using different proton hydration scales. For Rb and K, the gas-phase cluster free energies are closer to experiment for smaller cluster sizes while larger deviations exist for larger clusters. The MD results have generally smoother variations as a function of cluster size. The harmonic approximation results, however, has a significant change in slope when cluster size increases from 4 to 5. This change is likely because of the limitation of harmonic approximation and the insufficient sampling of all possible structures. This sharp change in slope is also reflected in the final hydration free energy results. For Rb, the MD value does not converge at n = 5, the largest cluster size for which inner-shell structure can be obtained. The harmonic approximation value converges faster than the MD value and falls within the range of the two experimental estimates, Marcus and Schmid values, that do not include surface potential.[73] However, this agreement is caused by an error cancelation, since the harmonic approximation is less accurate for the free energy of larger clusters. There are also difference between results using PCM or SMD

solvation models, although the difference is minimized at n = 3 or 4. According to MD results, the QCT hydration free energy is between experimental estimates with and without surface potential, whereas the prediction from harmonic approximation is consistent with experimental estimates without surface potential.

The results for K+ using harmonic approximation also show a sharp change in the slope of both gas-phase cluster free energies and hydration free energy. However, the hydration free energy does not converge with cluster size, but rather has a minimum at n = 3. This minimum free energy is close to the Marcus value. MD results have much better agreement with experimental gas-phase cluster free energies, and shows convergence in hydration free energy when combined with the SMD solvation model. Similar to the case of Rb+, the converged hydration free energy from MD/SMD is between experimental estimates with and without surface potential.

The calculation of F- hydration free energy is more challenging because of the increased tendency for covalent bonding. Harmonic approximation results have much better agreement with experimental gas-phase cluster free energies than MD results do. This may be attributed to the inclusion of quantum effect in the harmonic approximation. However, the hydration free energy predicted by harmonic approximation is much closer to Tissandier value rather than the values without surface potential. This is inconsistent with the results for Rb and K. Due to this inconsistency, the total hydration free energy for an counter-ion pair predicted by the harmonic approximation will be very different from experimental measurements. On the hand, the hydration free energy predicted by MD/SMD is consistently between Tissandier values and Schmid values.

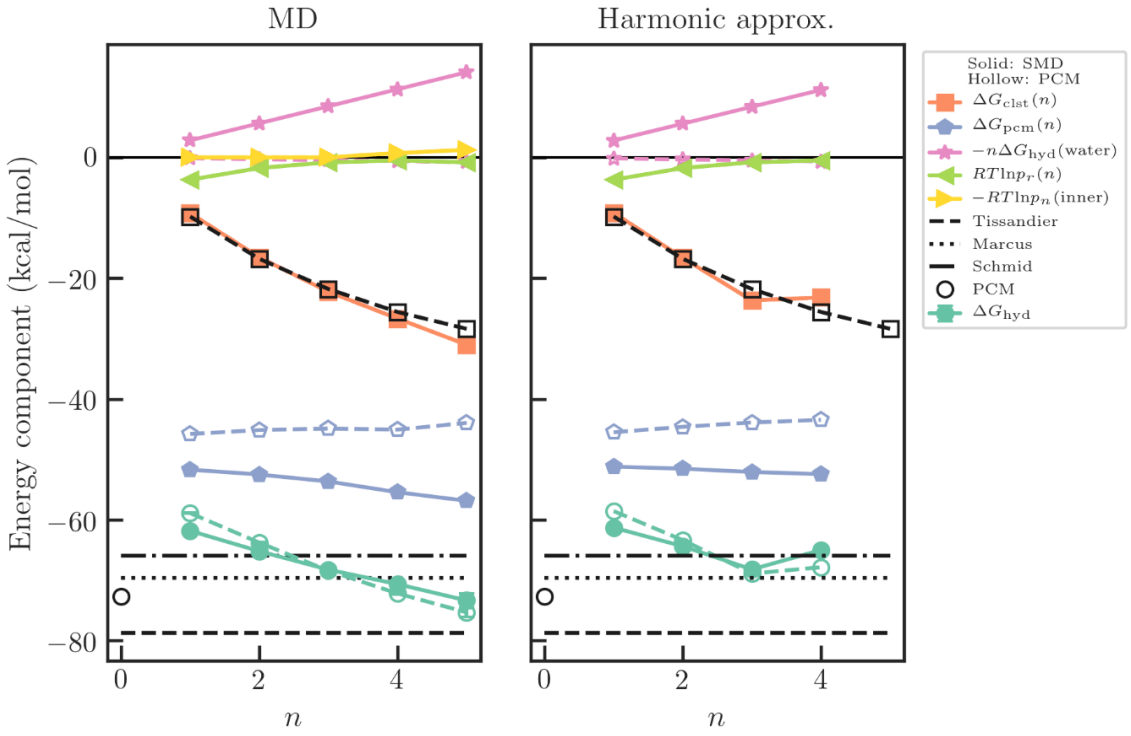


Figure 26: Rb+ hydration free energy calculated by QCT.

The QM solvation energy is calcualted by SMD (solid stars, pentagons and spheres) or PCM (hollow stars, pentagons and spheres).

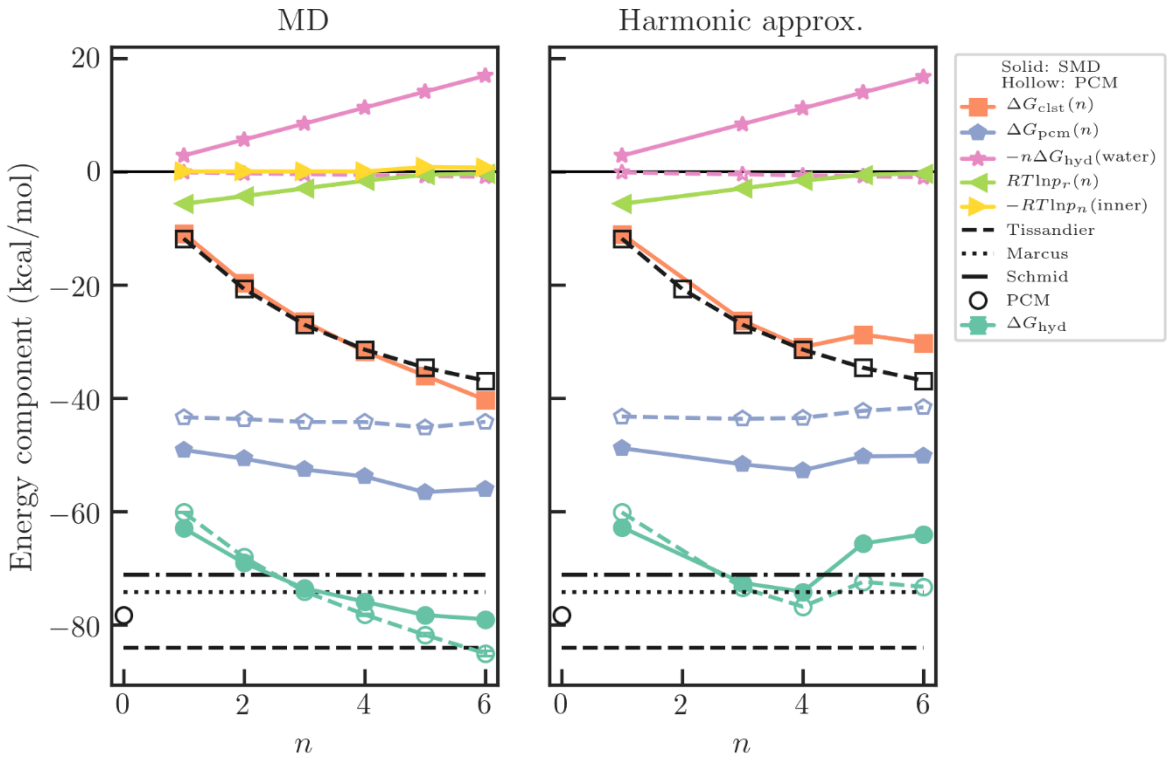


Figure 27: K+ hydration free energy calculated by QCT.

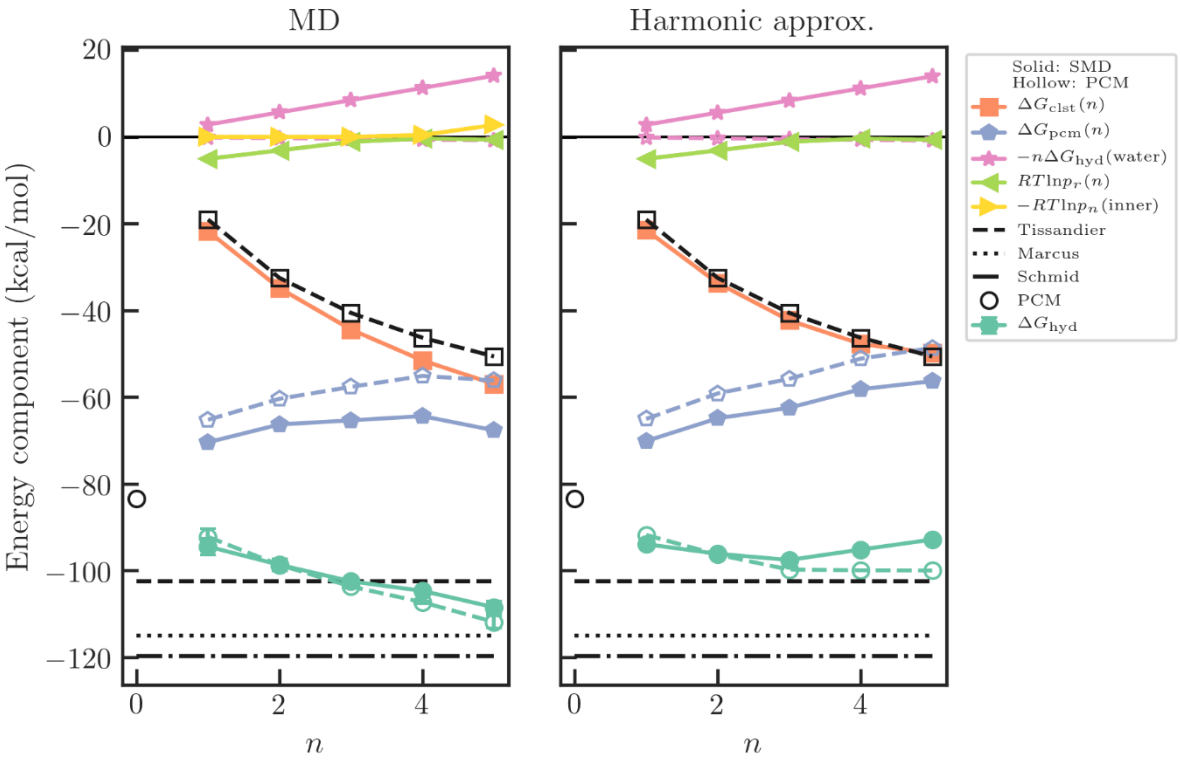


Figure 28: F- hydration free energy calculated by QCT.

4. SUMMARY

DFT methods were compared for their accuracy in predicting ion interactions, especially those with Rb and K. It was found that B97M-V and MN15, when combined with a sufficiently large basis set such as aug-cc-pVDZ, have very good accuracy. The DFT methods were compared to both high-level CCSD(T) calculations and experimental measurements of gas-phase interactions energies. However, there are large the variations between experimental techniques and the same techniques used by different groups. The experimental variations are much larger than the error of most of recent DFT methods. Therefore, currently gas-phase experimental data cannot be used to benchmark DFT methods. On the contrary, experimental techniques can be evaluated based on their agreement with state-of-the-art QM methods.

The hydration free energy of Rb, K and F ions was calculated by QCT using different sampling methods and solvation models. Generally, good agreement is found on the gas-phase cluster free energies for small cluster sizes. The hydration free energy is also within the range of experimental estimates. Despite the use of rigorous sampling method, the predictions still suffer from convergence issues, in addition to the uncertainties caused by the implicit solvation models. Therefore, it is unclear which experimental convention is more consistent with QCT.

5. METHODS

5.1 Free energy calculation through harmonic approximation

According to both the overall performance and the accuracy for cation-water dimers and water clusters in the GMTKN55 benchmark[152], the best metaGGA's are TPSS-D3, revTPSS-D3 and M06L-D3, the best hybrids are ωB97X-V and ωB97X-D3, and the best double-hybrids are DSD-BLYP-D3 and DSD-PBEP86-D3. None of the GGA's in that study have balanced performance. MN15 was chosen in our study based on its good performance for ion interactions in our benchmark. The basis sets are aug-cc-pVDZ for water, and def2TZVPP for ion. For the implicit solvation models PCM/SMD, the dielectric constant was set to eps=78.35, and default radii were used except for Rb (3.2 Å) and K (3.5 Å).

20 structures from AMOEBA MD simulation at 300 K were optimized by QM using ultrafine integration grid and any imaginary frequency was removed by further optimization. The lowest free energy structure for each cluster size n was chosen for the free energy calculation.

In $M^+(H_2O)_n$ cluster, the n water molecules are indistinguishable. The contribution of this multiplicity to free energy $-RT\ln n!$ is -3.9 kcal/mol for $n = 6$.

When the contribution of multiplicity is included, the agreement on K+ fee energy is better but the agreement on Rb+ is somewhat worse.

The cluster binding free energy and hydration free energy are calculated as follows. Assuming the n water molecules are indistinguishable,

$$\Delta G_{\text{clst}}(n) = G(\text{MW}_n) - G(\text{M} + n\text{W}) = -k_{\text{B}}T \ln \frac{Q(\text{MW}_n)}{Q(\text{M} + n\text{W})} \tag{5.2}$$

$$Q(\text{MW}_n) = \frac{1}{\lambda^{3N} n!} \int_{\text{MW}_n} \text{d}R_0 \text{d}R_1 \ldots \text{d}R_n e^{-\beta U} \tag{5.3}$$

$$Q(\text{M} + n\text{W}) = \frac{1}{\lambda^{3N} n!} \int_{\text{M}+n\text{W}} \text{d}R_0 \text{d}R_1 \ldots \text{d}R_n e^{-\beta U} = Q(\text{M}) \frac{1}{n!} Q(\text{W})^n \tag{5.4}$$

$$\Delta G_{\text{clst}}(n) = G(\text{MW}_n) - G(\text{M} + n\text{W}) = G(\text{MW}_n) - G(\text{M}) - nG(\text{W}) - k_{\text{B}}T \ln n! \tag{5.5}$$

where M and W stand for metal cation and water, $\text{M} + n\text{W}$ denotes non-interacting ion and n water molecules, $N = 3n + 1$ is the number of atoms, R_0 and $R_i\ (1 \le i \le n)$ indicates the positional degrees of freedom of cation and water, respectively. Note that assuming distinguishable water molecules will lead to the same result.

Let $Q\left(\text{MW}_n^{(0)}\right)$ denote the partition function from vibrational analysis of a single structure with distinguishable atoms, i.e. the output from Gaussian16. Then $Q(\text{MW}_n) = \frac{s}{\sigma} Q\left(\text{MW}_n^{(0)}\right)$, where σ is the rotational symmetry number and $s = 2$ for chiral structure and $s = 1$ otherwise. Similarly $Q(\text{W}) = \frac{1}{2} Q\left(\text{W}^{(0)}\right)$, since the rotational symmetry number of the water molecule is 2. In general, we have the following relationship,

$$\begin{aligned} Q(\text{MW}_1) &= \frac{1}{2} Q\left(\text{MW}_1^{(0)}\right) \\ Q(\text{MW}_2) &= Q\left(\text{MW}_2^{(0)}\right) \\ Q(\text{MW}_3) &= \frac{2}{3} Q\left(\text{MW}_3^{(0)}\right) \\ Q(\text{MW}_n) &= 2Q\left(\text{MW}_n^{(0)}\right) \text{ for } n \ge 4 \end{aligned} \tag{5.6}$$

$Q(\text{MW}_n) = 2Q\left(\text{MW}_n^{(0)}\right)$

5.2 Free energy calculation through AIMD simulation

Geometry optimization and vibrational analysis were used to obtain starting positions and velocities from the canonical ensemble at 300 K for AIMD simulations.[153]

The AIMD simulations were conducted under the NVT ensemble at 300 K by using the canonical sampling via velocity rescale thermostat with a time constant of 20 fs. The integration time step was 0.5 fs. The center of mass velocity was removed.

A non-periodic cubic box of $(20\ \text{Å})^3$ was used and the clusters were placed at the center of the box. Gaussian and plane wave (GPW) method with MOLOPT-TZV2P and a plane wave basis set with a cutoff of 400 Ry was used AIMD. For geometry optimization and vibrational analysis, a larger cutoff of 640 Ry was used to improve the convergence.

The structures from AIMD were analyzed using Gaussian 16. The free energy was calculated using the double-decoupling method.

In AIMD, the cluster free energy is calculated by

$$\Delta G_{\text{clst}}(n) = \sum_{i=1}^{n} [\Delta G_{\text{clst}}(i) - \Delta G_{\text{clst}}(i-1)] \tag{5.7}$$

$$\begin{aligned}\Delta G_{\text{clst}}(n) - \Delta G_{\text{clst}}(n-1) \\ = G(\text{MW}_n) - G(\text{MW}_{n-1}) - G(\text{M} + n\text{W}) + G(\text{M} + (n-1)\text{W})\end{aligned} \tag{5.8}$$

$$\begin{aligned}G(\text{MW}_n) - G(\text{MW}_{n-1}) &= k_{\text{B}}T \ln\langle e^{-[U^{\text{MW}_{n-1}} - U^{\text{MW}_n}]/k_{\text{B}}T}\rangle_{\text{MW}_n} \\ &= -k_{\text{B}}T \ln\langle e^{-[U^{\text{MW}_n} - U^{\text{MW}_{n-1}}]/k_{\text{B}}T}\rangle_{\text{MW}_{n-1}}\end{aligned} \tag{5.9}$$

$$G(\text{M} + n\text{W}) - G(\text{M} + (n-1)\text{W}) = k_{\text{B}}T \ln\langle e^{U^{\text{W}}/k_{\text{B}}T}\rangle_{\text{W}} \tag{5.10}$$

where U^{MW_n} means the potential energy calculated by the Hamiltonian of the MW_n system, $\langle \ldots \rangle_{\text{MW}_n}$ means the average calculated over the canonical ensemble of MW_n. The $-k_{\text{B}}T \ln n!$ term is implicitly included. In principle, the free energy change $G(\text{MW}_n) - G(\text{MW}_{n-1})$ can be evaluated by either deleting or inserting one water molecule, the latter of which is also known as Widom insertion. However, the former approach does not account for the entropy change of the water molecule moving from

standard concentration to a confined volume, which varies among different cluster sizes. It is noted that Muralidharan et al.[151] attempted to use the same MW_n trajectory to evaluate $G(\mathrm{MW}_n) - G(\mathrm{MW}_{n-1})$ and $G(\mathrm{MW}_1) - G(\mathrm{M})$ and thus avoid direct evaluation of the entropy term. Despite a minor mistake in the derivation, this scheme also works in principle. Below the correct derivation is provided

$$[G(\mathrm{MW}_n) - G(\mathrm{MW}_{n-1})] - [G(\mathrm{MW}_1) - G(\mathrm{M})] = k_\mathrm{B}T \ln \frac{\langle e^{-\Delta_a/k_\mathrm{B}T}\rangle_{\mathrm{MW}_n} \langle e^{-\Delta_b/k_\mathrm{B}T}\rangle_{\mathrm{MW}_n}}{\langle e^{-\Delta_c/k_\mathrm{B}T}\rangle_{\mathrm{MW}_n}} \quad (5.11)$$

where $\Delta_a = U^{\mathrm{MW}_{n-1}} - U^{\mathrm{MW}_n}, \Delta_b = U^{\mathrm{MW}_1} - U^{\mathrm{MW}_n}, \Delta_c = U^{\mathrm{M}} - U^{\mathrm{MW}_n}$. Choosing the same water molecule for Δ_a and Δ_b, or evaluating the ensemble average using each of the n water molecules, can reduce the error due to insufficient sampling of AIMD. However, Δ_b and Δ_c are more difficult to sample than Δ_a, which is the quantity of interest for the deletion method. The errors are illustrated in Figure 29.

On the other hand, the Widom insertion approach using QM energy calculation is very expensive, although it can be partly alleviated by pre-sampling with a faster method such as force field calculation. Here we take a different approach where the standard binding free energy is calculated rigorously by using the double-decoupling method with a force field, and the free energy difference between force field and QM is also calculated by free energy perturbation. The AMOEBA+ force field[47,62] was used since it is accurate for water clusters and ion-water interactions and it reproduces the equilibrium angle of water in both gas phase and liquid phase.

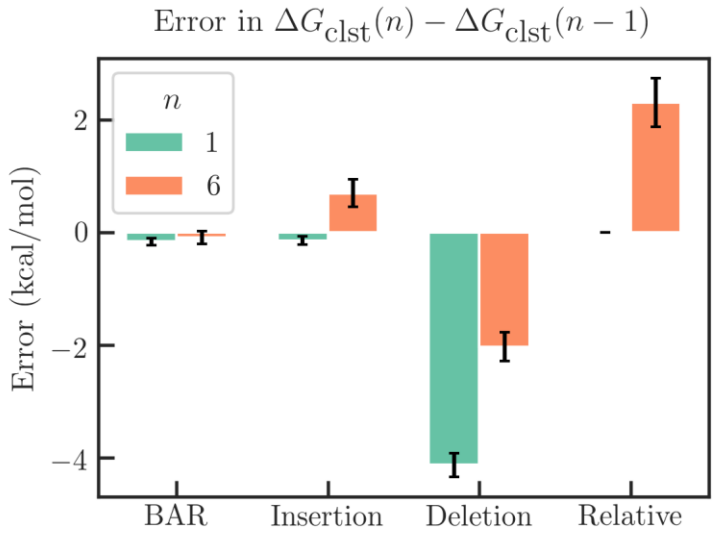


Figure 29: Errors of free energy calculation methods estimated from AMOEBA simulation of K water clusters.

The deletion and the relative method used the trajectory of MW_n, the insertion method used the trajectory of MW_{n-1}, while the BAR method combined data from both trajectories. The relative method is a re-derivation of Muralidharan's method. Each method used a total of 10000 structures. The reference value was calculated by the alchemical free energy method with 19 intermediate steps and 2000 structures in forward/backward direction of each step. The insertion process consisted of randomly placing a water molecule with ion-oxygen distance between 2.5 and 3.7 Å, and the entropy change from standard concentration to the concentration corresponding to this shell was added to the free energy.

The cluster binding free energy is calculated as follows.

$$\begin{aligned}\Delta G^{\mathrm{QM}}_{\mathrm{clst}}(n) - \Delta G^{\mathrm{QM}}_{\mathrm{clst}}(n-1) &= \left[\Delta G^{\mathrm{MM}}_{\mathrm{clst}}(n) - \Delta G^{\mathrm{MM}}_{\mathrm{clst}}(n-1)\right] + \left[G^{\mathrm{QM}}(\mathrm{MW}_n) - G^{\mathrm{MM}}(\mathrm{MW}_n)\right] \\ &\quad - \left[G^{\mathrm{QM}}(\mathrm{MW}_{n-1}) - G^{\mathrm{MM}}(\mathrm{MW}_{n-1})\right] - \left[G^{\mathrm{QM}}(\mathrm{W}) - G^{\mathrm{MM}}(\mathrm{W})\right]\end{aligned} \tag{5.12}$$

where the first term on R.H.S. is calculated by BAR with alchemical transition in 19 steps, and the remaining three terms are calculated by BAR with one step.

AMOEBA+ simulations

The AMOEBA+ simulations of the non-periodical systems were performed at 300 K by using the canonical sampling via velocity rescale thermostat. The RESPA integrator

with a time step of 0.5 fs was used. The small time-step was chosen because of the large temperature fluctuations expected for the system size. For each simulation, the length was 2 ns and 2000 frames were saved for analysis. The OPT4 algorithm was used to solve the polarization. Ion parameters were obtained by fitting to ion-water dimer SAPT2+ energy component data and ion-water cluster DFT energy data. A development version of TINKER 8 that implements AMOEBA+ with geometry-dependent charge-flux, available at https://github.com/TinkerTools/, was used for the simulations.

In the alchemical free energy calculation, 20 alchemical states were used for the binding free energy of one water. The vdW interaction was first turned on in 9 steps before the electrostatics interaction was turned on in another 10 steps. A flat-bottom distance restraint between the ion and the water molecule was imposed when there is no or weak interaction. The free energy difference of water moving from standard gas-phase concentration at 1 atm to the restrained volume was calculated analytically.

Chapter 6: Ion binding thermodynamics in potassium channels

1. INTRODUCTION

Potassium channels are present in all mammalian cells. They are responsible for maintaining resting potential, regulating action potential, hormone secretion and signaling. Potassium channels are characterized by a combination of high conduction rate and high selectivity for K+ against Na+.[154] Numerous studies have been devoted to understand the ion conduction mechanism in Potassium channels.[154-160] The crystal structure of KcsA reported by Zhou et al. provides a clue of the conduction mechanism.[161] In the crystal structure, there are four K+ binding sites arranged in a single file and the distance between neighboring sites are only 3-3.5 Å. Since ions of like charge strongly repel each other, it was hypothesized that the four K+ sites are not all occupied simultaneously. Rather, there are two alternating states, with K+ occupying sites 1,3 or 2,4, and the overlap between the two states lead to the four K+ ions in the crystal structure.[161,162] This hypothesis is also known as "soft knock-on" mechanism because K+ ions are separated by water molecules. The soft knock-on mechanism has served as a starting point for ion-binding models and MD simulations for two decades. The soft knock-on mechanism is also confirmed by MD simulations[163] of ion permeation and free energy calculation.[164,165] Recently, this hypothesis was challenged by (i) a reanalysis of crystallography data showing the occupancy of each of the four binding sites is near unity,[166] and (ii) direct MD simulations which shows that most of the ion permeation events are not accompanied by water.[166,167] Nevertheless, the four ion states in the crystal structure were rarely observed in the simulations, apparently contradicting the crystallography data;[166,167] the four-ion configuration was only observed in 20-ns simulations at low temperature,[166] which could possibly result from the slow dynamics at

low temperature. Simulation results also depended on the force field parameters, with some force fields showing higher ratios of co-permeating water.[167] Free energy simulation found that a water-free mechanism and the soft knock-on mechanism are both possible due to the similar conduction barriers.[168,169]

Experimental evidence supporting either direct or soft knock-on have been reported. The ion occupancy in NaK2K was determined to be 4 by anomalous X-ray diffraction.[170] Study of the NaK2K channel under physiological conditions by using solid-state NMR shows that water does not interact with the SF.[171] Cuello and coworkers[172] identified a water-ion-water-ion state in the G77A mutant of KcsA. The bridging carbonyl groups between S2 and S3 are absent in the mutant. Despite the difference in binding sites, ion conduction property and selectivity were maintained. In addition, the ion binding properties of the wildtype and the mutant are virtually the same. This result supports the soft knock-on mechanism as a possible pathway for selective conduction of K+.

Currently, the ion conduction mechanism is still debatable because of the contradictory or ambiguous evidence. An accurate ionic model is required to determine the preference for the two competing mechanisms through MD simulations, as an imbalance in the interactions could shift the relative stability. Most of the previous simulations employed fixed-charge force fields.[155] It has been shown that fixed-charge force fields yield unrealistic ion-binding free energies[173] and the wrong order for Mg2+/Ca2+ selectivity in calcium-binding proteins,[1,106] while polarizable force fields dramatically improve the accuracy for these systems.[1,48,173-175] The SF in potassium channels binds multiple K+ ions at a time, and therefore the polarization effect is also expected to be significant. In this work, we tried to provide a clue of the conduction mechanism by studying the thermodynamics stabilities of different states by using

polarizable force fields. Unlike in previous simulations using fixed-charge force field, the four-ion states were dominant in our simulations. Therefore, the four-ion state is possibly an important intermediate state in the conduction mechanism. Furthermore, the results show that the KcsA-G77A mutant prefers the soft-knock mechanism due to the altered ion-binding properties.

2. RESULTS

2.1 Ion binding in KcsA.

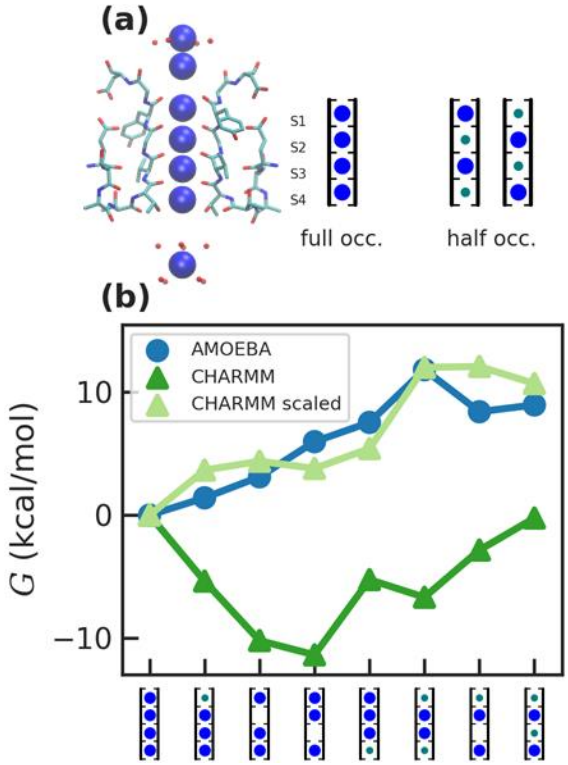


Figure 30: Free energies of different states in the selectivity filter of KcsA.

(A) Crystal structures and hypotheses of the ion configurations of KcsA. (B) Standard free energies predicted by free energy calculation using different force fields. Blue and green dots represent K+ ion and water molecule. The statistical errors are below 1 kcal/mol.

We determined the standard free energy of the different structures in proposed conduction mechanisms through MD free energy calculation methods. The AMOEBA force field was chosen for the simulations, since it reproduces gas-phase interaction energy,[176] solvation free energy and binding free energies for both monovalent and divalent ions.[1] Based on the electron densities in the X-ray crystal structures of KcsA and NaK2K, all four sites are occupied simultaneously by ions.[166,170] An alternative interpretation known as the soft knock-on mechanism states that ions in the SF are always separated by a water molecule and only half of the binding sites are occupied at the same time (Figure 30).[162,172] In the direct knock-on mechanism observed in previous work, the most stable conformations contain 2-3 ions in the SF and water rarely co-permeate with ions. To determine which mechanism has more favorable ion configurations, we calculated the standard free energies of various conformations by AMOEBA as shown in Figure 30. Although the four-ion state was considered unstable due to strong repulsion between the ions, it was predicted to be most stable under standard condition. The standard free energy of some three-ion states is slightly higher than that of the four-ion state; under physiological K+ concentrations of 5-150 mM their relative stabilities could be flipped. Experimentally, different ion occupancies were observed for K+ concentration below and above 5 mM.[161] The results of CHARMM favored reduced occupancy of 2 and 3, which is in stark contrast with the AMOEBA results. Previous simulations with CHARMM and AMBER force fields also led to reduced occupancy.[166,167] We also conducted simulations using CHARMM with the charges of all ions and charged residues scaled by 0.7 to model polarization effectively. The results also show that the four-ion state is the most stable. Other factors, such as temperature (298 vs. 200 K) and lipid environment were also studied, which do not qualitatively affect the trends of stabilities. The results are insensitive to force field parameters. The agreement between the two

different approaches to modeling polarization effect suggests that the reduced ion occupancy in previous simulations is an artifact of the fixed-charge force fields used in those studies.

2.2 Ion binding in KcsA-G77A.

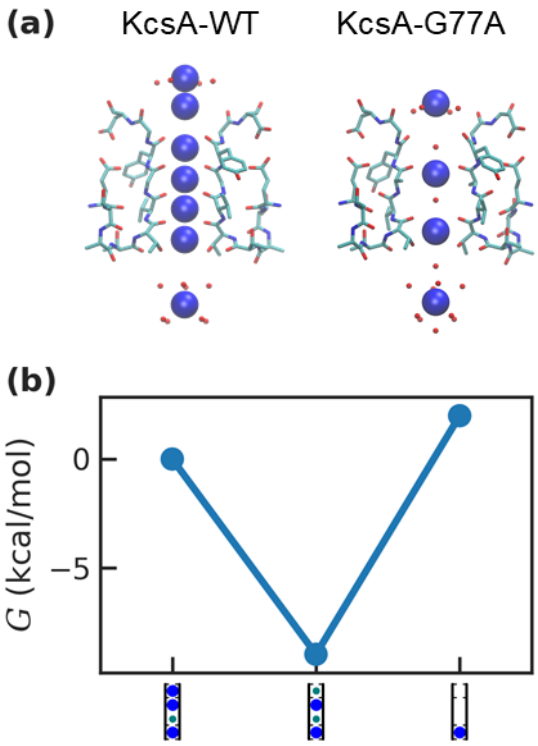


Figure 31: Free energies of different states in the selectivity filter of G77A mutant of KcsA.

(A) Structures of KcsA-WT and KcsA-G77A (PDB: 1k4c and 6nfu); (B) Standard free energies predicted by free energy calculation using AMOEBA. Blue and green dots represent K+ ion and water molecule. The statistical errors are below 1 kcal/mol.

One of the recent studies supporting the soft knock-on mechanism is the mutagenesis study on the G77A and G77C mutants of KcsA. In the crystal structure of KcsA-G77A, the water-K+-water-K+ state was observed. (Figure 31)[172] Due to the rotation of the carbonyl groups between S2 and S3, there is no ion binding at S3. KcsA-G77A has smaller conduction rate than the wildtype, but it still selectively conducts K+. In addition, the ion binding affinities of the wildtype and the mutant are virtually the same.[172] The similar ion conduction and binding properties between KcsA-G77A and KcsA-WT raised the possibility that KcsA-WT could also have the water-separated ion configuration.

As our free energy calculations have predicted KcsA-WT to have an occupancy of 4, we next applied the same approach to understand the different ion configuration in KcsA-G77A. The results show that water-K+-water-K+ has a lower standard free energy than other possible 3-ion and 1-ion states. (Figure 31). Compared to the wildtype, the binding sites S1 and S3 both lose K+ ion, although the mutation is on S2 and S3. Cuello and coworkers hypothesized that the loss of K+ at S1 and S3 is an indication of the soft knock-on mechanism.[172,177] Here we found that the coupling of ion binding at the two sites is due to strong interaction between ions rather than the existence of two possible configurations: the upward shift of the ion at S2 makes ion binding at S1 unstable.

2.3 Ion binding free energies.

As a further validation, we compare the ion binding affinities from isothermal titration experiments and our simulations (Table 9). Lockless and coworkers[178] showed that the ion binding of K+ in common potassium channels in ion binding experiments is the competitive binding. The intrinsic binding affinity of K+ from the competitive

binding model is ~0.1 mM. The selectivity factor between Na+ and K+ is determined to be 10^2~10^4 for different mutants. Thermal denaturation titration experiments found a binding affinity of ~2 μM for K+.[179] In our simulations, there are three K+ binding events for the conductive conformation of KcsA, leading from 1 K+ to 4 K+ ions. The first and the third binding events are likely either too strong or too weak to be measured by experiments; the second binding was predicted to have similar ΔG as the experimental value (-6.8 kcal/mol vs. -5.6 kcal/mol).[178] The binding in the collapsed SF conformation was predicted to have a ΔG -4.9 kcal/mol, compared to the experimental value of -4.7 kcal/mol. The predicted relative binding free energies between Na+ and K+ were also in good agreement with experimetns.[178]

		AMOEBA		ITC[f]	
Protein	SF conformation	K+	Na+→K+	K+	Na+→K+
KcsA-WT	collapsed	-4.9 ± 0.7	-1.6 ± 0.3[d]	-4.7	-2.1
KcsA-WT	conductive	-24.8 ± 0.6[a] -6.8 ± 0.6[b] -1.4 ± 0.6[c]	-2.3 ± 0.2[e]	-5.6	-2.9
KcsA-G77A	conductive	-6.0 ± 0.5	-3.3 ± 0.2[d]		

[a] ΔG between 1 K+ and 2 K+ states;
[b] ΔG between 2 K+ and 3 K+ states;
[c] ΔG between 3 K+ and 4 K+ states;
[d] ΔG between 2 Na+ and 1 Na+/1 K+ states;
[e] ΔG between 1 Na+/3 K+ and 4 K+ states.
[f] Liu et al.[178]

Table 9: Comparison between calculated and experimental ion binding free energies (kcal/mol).

The almost identical ion binding affinities between the wildtype and G77A mutants were used to support the soft knock-on mechanism. If the wild type and the

mutant have two different conduction mechanism, the difference should be reflected in the ion binding affinity. However, the apparent binding affinities of K+ for KcsA-WT and KcsA-G77A are 0.30 and 0.29 mM, respectively, which suggest that they have the same ion configuration. However, KcsA-WT and KcsA-G77A have different conformational states associated with the measured binding event, so it is difficult to interpret the binding affinity data.[178,180] MD simulations with AMOEBA showed that the binding free energy of K+ in two conformations of the wildtype and in the G77A mutant are close to each other within the errors of simulations. (Table 9). The results imply that the similarity between the binding free energies for the wildtype and the mutant may not be connected to the same ion configuration.

2.4 Ion conduction barriers of different mechanisms.

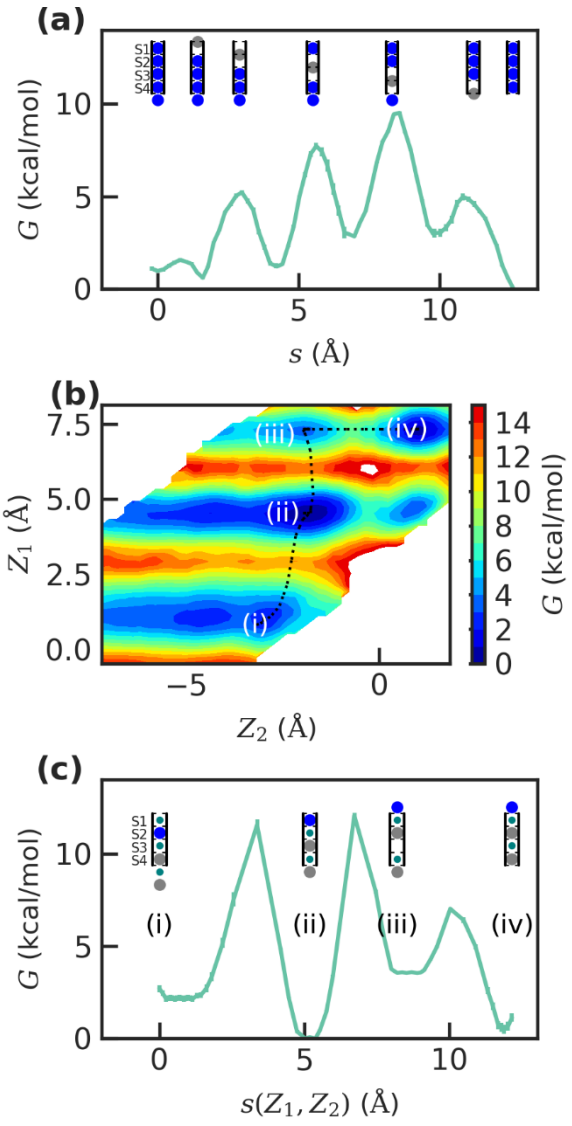


Figure 32: Ion conduction barriers for KcsA.

(A) Conduction barrier for the single-vacancy mechanism; (B) and (C) Conduction barrier for the soft knock-on mechanism. The two-dimensional free energy is expressed as a function of two ion coordinates. The lowest energy path in the two-dimensional map in (B) is projected onto one-dimensional free energy in (C). The restrained ions are shown in gray, free ions are shown in blue and water molecules are shown in green.

Since the free energy of the two-ion states in KcsA is ~5 kcal/mol higher than that of the four-ion configuration, the 2,4-bound configuration is unlikely a dominant intermediate state during ion conduction. To interrogate this hypothesis, we performed umbrella sampling of the conduction pathways[181] to compare the soft and direct knock-on and the single-vacancy mechanisms. As shown in Figure 32, the four-ion states were found to be more stable compared to one vacancy at either S1 or S4, which agrees with the free energy calculation of KcsA solvated in water (Figure 30). The highest barrier is about 10 kcal/mol, while the individual barriers are at most 6.5 kcal/mol. (Figure 32). The calculations were performed using a partially open conformation, which does not allow additional ions to come into or out of the cavity. The free energy landscape could be changed if a different conformation is adopted. Compared to previous work, the barriers are much larger.[164,168] This could be explained by force field error or sampling/conformation. Relatively speaking, the barrier of the single-vacancy mechanism is still smaller than the barrier of the soft knock-on mechanism (Figure 32). These results provide additional support for the direct knock-on mechanism.

3. DISCUSSION

One argument for the soft knock-on mechanism was that K+ ions are highly repulsive to each other and thus they cannot form direct ion pairs.[162] The direct knock-on mechanism of de Groot and coworkers[166,167] showed that the strong repulsion is necessary in reducing the conduction barriers. As we have demonstrated in our free energy calculations, the strong repulsion between K+ ions is an artifact of the force fields employed in previous simulations and it is incompatible with crystallography data of ion occupancy. Polarization effect makes K+ ions much less repulsive, which stabilizes the

fully occupied configuration. The high K+ occupancy accommodated by the chemical and electrostatic environment of the SF may play a role in the high conduction rate and selectivity.

Some mutagenesis studies have also been used to support the soft knock-on mechanism. Using mutagenesis to modify or remove some of the binding sites has provided many novel insights.[177,180,182-184] For example, studies on the NaK channel and its mutants showed that the 4-site channels are K+ selective while the 3-site and 2-site channels are nonselective.[185] The KcsA T75A mutant has inverted allosteric coupling between activation and inactivation gate, i.e. the closed inactivation gate is coupled to the collapsed conformation of the SF, which reveals the role of T75 in the allosteric coupling.[186] However, one frequent assumption is the independence of K+ binding at each site. Correlation between the occupancies at nonadjacent sites have been interpreted as evidence for multiple alternating configurations. Here we showed that in the case of KcsA G77A mutant, coupling between both adjacent and nonadjacent sites occurs because of strong ion-ion interactions rather than the coexistence of multiple states.

Other experiments supporting the soft knock-on mechanism includes the water/ion co-translocation ratio derived from streaming potential measurements[187] and two-dimensional infrared (2D IR) spectra of the SF.[188] As pointed out by de Groot and coworkers,[166,167] the streaming potential measurements were carried out under a different condition from that of ion conduction experiments, and the 2D IR spectra are compatible with different conformations. Due to the limit of time scale, our simulations cannot directly explain these observations.

Our simulations with polarizable force fields reconcile the discrepancy between the ion conduction mechanisms and crystallography studies on ion occupancy. Understanding the ion occupancy is important because it likely affects the structure and

dynamics of the SF as well as the inactivation mechanism. MD simulations have been widely used to study ion channels.[189-192] Since polarizable MD simulations have more realistic ion occupancies than previous fixed-charge simulations, we expect that future work with polarizable force fields will shed more light on these problems.

4. METHODS

Force field validation. The AMOEBA force fields were compared to high-level QM CCSD(T)/CBS and MP2/CBS methods on ion-water, ion-amino acid interactions. Modifications on the carbonyl multipole and polarization parameters were explored. The performance of AMOEBA is on par with those of DFT methods such as B3LYP-D3BJ. The default AMOEBA protein[193] and ion parameters[68] were used, except modifications on the carbonyl polarization parameters as described Liu et al.[194] The lipid parameters were taken from Li and coworkers.[195]

Crystal structures of KcsA in different conformations (1k4c/d, 3fb6, 6nfu) were used as a starting point to build the model systems. The partially open conformation was used to study conduction barrier, while other conformations were used for equilibrium free energy calculation. The protonation states of titratable residues were assigned based on pH=7 except that Glu71 is protonated and forms a hydrogen bond with Asp80.[167] The proteins were either solvated in water or lipid bilayers. The box size was $(7.2\ \text{nm})^3$ for the water box and $7.6\times7.6\times9.2\ \text{nm}^3$ for the lipid box. CHARMM-GUI[196] was used to build the systems and generate the initial simulation parameters. Ions were added to reach a final concentration of 0.15 M NaCl and 0.4 M KCl for the water box and lipid box, respectively.

Position restraints on α-C atoms were used to maintain the protein conformation. [175] No restraints were exerted on the SF. The force constants were 5.0 kcal/mol/Å and the

restraint is turned on outside a radius of 2.0 Å. GROMACS 2018[197] was used for fixed-charge force fields and Tinker-OpenMM[102,198] was used for AMOEBA simulations.

The binding free energies were calculated by using the double decoupling method.[58] To reduce the finite-size effect in ionic binding free energy, the ion solvation free energy was calculated with the same system used for ion-protein binding.

The conduction barriers were calculated by using umbrella sampling and analyzed by using weighed histogram analysis method. The WHAM code of Alan Grossfield[199] was used to reconstruct the PMF from simulation trajectories.

Chapter 7: Hybridization stability of modified nucleic acids

Note: this Chapter is based on a published paper[200] for which I conducted the experiment, analyzed the data and wrote the paper, Jing, Z., Qi, R., Thibonnier, M. & Ren, P. Molecular Dynamics Study of the Hybridization between RNA and Modified Oligonucleotides. J. Chem. Theory Comput. 15, 6422-6432 (2019).

1. INTRODUCTION

Gene therapy is very promising for treating various diseases, especially ones without druggable targets.[201] The target in gene therapy is mostly determined by the nucleobase sequence, which allows for rapid development of the therapeutic agents.[202] The action mechanisms of gene therapy include enhancement or repression of messenger RNAs and microRNAs. The therapeutic agents can be classified according the action mechanism, including microRNA (miRNA) mimics, anti-microRNAs (antagomirs), among others.[201] Several candidates of gene therapy, including siRNA and miRNA, have either been approved by the FDA[203] or in clinical trials. Development of other types of gene therapy are still in the early stage.[201]

MiRNAs are small non-coding RNAs that plays a role in gene regulation by binding to mRNAs.[204,205] It is estimated that more than 50% of genes are regulated by miRNAs.[206,207] Therefore, it is not surprising that miRNAs are related to many diseases including cancer, neurological diseases, cardiovascular diseases and metabolic disorders. MiRNA also holds great therapeutic potential because each miRNA targets a network of many genes.[206,207] Therapeutics targeting miRNAs can either restore or inhibit the activity of miRNA.[206,208] The oligonucleotides that inhibit miRNA is called antagomirs. In a

recent work, we reported the potential of miRNA for treating obesity by modulating brown adipogenesis.[209-212]

The major challenge for gene therapy is the drug delivery. Chemical modifications have been introduced to improve the pharmacokinetic properties of oligonucleotides. In addition, the chemical modification can also improve the stability against nuclease.[202] Natural modifications of RNA and DNA exist.[213] The 2'-O-methylation was found to improve affinity and nuclease stability.[202] Synthetic modification was also introduced to increases nuclease resistance, such as phosphorothioate (PS).[202] Each PS modification has two stereoisomers, R and S (denoted Rp and Sp hereafter). Each isomer has a different biochemical and physical properties.[214-216] Other modifications that improve binding or pharmacokinetics include locked nucleic acid (LNA) and peptide nucleic acids (PNAs).[202,217-219] Structures of representative modifications are shown in Figure 1.

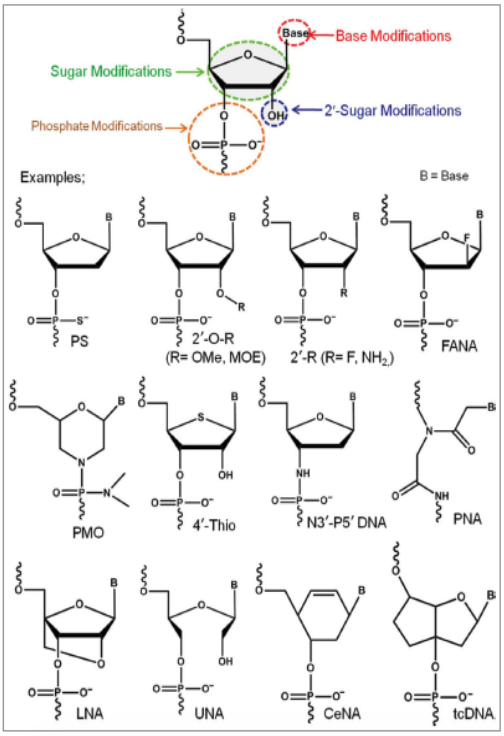


Figure 33: Structures of chemically modified nucleic acids.

The figure was taken from Ref.[200] with permission. Copyright of the American Chemical Society.

Molecular dynamics (MD) simulations and other computational techniques have proven useful for the study of modified nucleic acid.[220-225] Hartmann et al. studied sequence and chirality effects on the structural and energetic properties on PS hybrids.[222,226-228] Using QM calculations, Zhang et al. showed that Rp destabilizes the duplex because of the unfavorable repulsion between backbone and sugar.[227,228] A MD study by Mukherjee et al. suggest that the two PS isomers affect the minor groove differently.[226] MD simulations were also used to investigate the conformational preferences of LNA.[229-233] The conformational analyses showed that distance between phosphate in the two strands increases/decreases in DNA-RNA hydbrid/RNA duplex

upon the LNA modification.[229] Xu et al.[232] showed that the free energy change due to LNA modification can be efficiently calculated by FEP. MD simulations were used to study the conformations of PNA.[220,221,234,235] In the simulations, PNA has more dynamic backbones and the global helical structures are different from canonical DNA or RNA structures.[220,221] Verona et al. used enhanced sampling method to extensively sample the PNA structural preference.[234]

Although MD has been used in numerous studies[236-239] of nucleic acids and provided many insight, the short-coming of traditional force fields were recognized recently.[16,240,241] For example, simulations using earlies force fields could not reproduce the conformations of RNA tetranucleotides determined by NMR experiments.[241] The same problem was identified for PNA as well.[234] These short-comings illustrate the difficulty in modeling highly charged molecules such as nucleic acids.[17,242] Polarizable force fields is promising for the study of nucleic acids because of their better models for polarization and electrostatics.[56,243,244] In a previous work, we showed that the AMOEBA force field achieved excellent accuracy for protein-phosphate binding and protein-calcium binding.[104,245-249] The AMOEBA nucleic acids force fields also achieved state-of-the-art accuracy for both gas-phase and condensed-phase calculations.[243,244] Therefore, AMOEBA should be promising for the study of modified nucleic acids.

In this study, we developed the AMOEBA parameters for several modified nucleic acids and then use the force fields to investigate the thermodynamics of the hybridization. We will use FEP and BAR to calculate the hybridization free energies. Although BAR is a popular method for protein-ligand binding, it has not been used for the hybridization free energy. This may be because of the sampling and force field issues for nucleic acids. We will first compare our predictions to previous experimental data on PS, and then study different modifications of the miR-22 antagomirs.

2. METHOD

2.1 Force field.

DNA and RNA parameters were taken from our previous work and can be found from the Tinker distribution.[56,246] Parameters for the chemical modifications were developed following standard protocol.[55,248,250,251] The PS vdW parameters were determined by using MP2/CBS data on the PS-water dimer interactions energy. The PS torsion parameters were determined by using QM ωB97x-D/6-311++G(2d,2p)/PCM energy scan and AMOEBA/GK calculation. In the torsion fitting, all energies are calculated at the QM geometry.

2.2 System setup.

Two systems with experimental data were used for the validation: PS-6, which is 6-mer DNA/RNA hybrids or DNA duplexes with PS; PS-10 (PDB ID: 5J3F, 5J3I), which is 10-mer DNA duplex with PS.[215,252] The miR-22-3p systems were also set up with 8 to 12 nucleotides. PS, LNA, PS-LNA and PNA modified systems were constructed. For systems without crystal structures, a website server based on Amber was used to constructure the canonical structure (http://structure.usc.edu/make-na/server.html). Then the modifications were added. Na+ ions were added sequentially according to the ESP to neutralize the nucleic acids. Then 0.15M NaCl were added.

2.3 MD simulations

The initial structures were first minimized, then equilibrated with increasing temperature from 150 K to 298 K and then the density was relaxed in a total of 1.5 ns for each system. Production simulation were at constant temperature 298 K and volume.

All molecular dynamics simulations were run using the Tinker-OpenMM program[102] on GPU with a RESPA integrator,[253] Bussi thermostat,[254] Monte Carlo barostat, and 3.0 fs time step with hydrogen-mass repartition (heavy-hydrogen keyword). The cutoff distances for vdW and real-space electrostatics were 12 and 7 Å, respectively, while the long-range electrostatics were calculated by PME.

2.4 Hybridization free energy

The effect of the modification on hybridization was calculated by FEP and double decoupling.[102,250]

$$\Delta\Delta G_{\text{hyb}} = \Delta\Delta G_{\text{complex}} - \Delta\Delta G_{\text{solv}} \quad (7.1)$$

where $\Delta\Delta G_{\text{complex}}$ and $\Delta\Delta G_{\text{solv}}$ are the duplex and single-strand relative binding. Alchemical states were designed to allow for a smooth transition for each system. Detailed methodology can be found in previous papers.[255]

Since PS and LNA shared the same backbone with canonical DNA, a simple single-topology method was used for the FEP.[232] The transition from DNA to PS involves very small change, so 15~20 states were used for the PS systems. Systems with more PS modifications used a larger number of states. The transition from DNA to LNA involves a larger change than from DNA to PS. Also, the number of bonds is changed after the transition, which asks for more careful selection of the transition pathway. After some initial testing, the schedule for the transition of the C6'-O2' connection was determined as follows: (1) 8 steps for geometric sequences of bond force constant k_b increasing from 1e-4 to 0.5 kcal/mol/Å^2 and b_0 decreasing to 1.4565 Å. This ensures that the bond lengths change slower when the force constant is large. At the same time, the vdW parameters for the additional hydrogen was changed from 0 to 0.8. (2) 8 steps for the exponential change in the force constant from 0.5 kcal/mol/Å^2 to the final value accompanied by a change in

vdW from 0.8 to 1.0. (3) 16 steps for the exponential change in the force constant of the additional angles from to the final value accompanied by linear change in the angle parameter. (4) 14 steps for the electrostatics λ and other parameters. PS-LNA modifications were simulated in two stages, DNA to LNA and then to PS-LNA.

PNA has a different backbone from DNA, so FEP will be very inefficient. Therefore, the dependency on sequence length were studied. In each step, one nucleotide was grown using standard FEP technique. The schedule for alchemical transition was similar the our previous work.[249]

3. RESULTS AND DISCUSSIONS

3.1 Validation of PS simulations

Our simulation parameters were first validated by comparing to experimental thermodynamics on PS. We chose PS for validation because PS is relatively simple for simulations and there are plenty of experimental data. The crystal structures are shown in Figure 34.

The relative free energy from simulations and the Tm change from experiments are compared (Table 10). In the experiments, different trends were found for DNA/DNA and DNA/RNA. Rp stabilizes DNA/DNA and destabilizes DNA/RNA; Sp stabilizes DNA/RNA and destabilizes DNA/DNA. Similar trends were also observed in other work[214,216] except for some special sequences.[214] The MD simulations also suggest that Rp destabilizes DNA/RNA and Sp destabilizes DNA/DNA. The trends for DNA/DNA and DNA/RNA were both captured by the free energy results, indicating good reliability of our protocol.

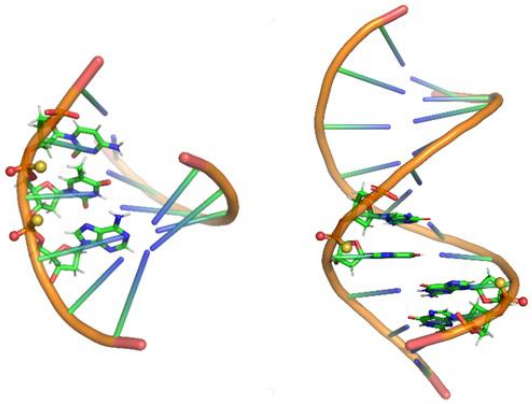


Figure 34: Crystal structures of stereodefined phosphorothioate DNA hybrids.

Hexamer [Rp,Rp]-DNA/RNA (left) and decamer Rp-DNA/Rp-DNA systems (right). PDB code: 5J3I.

Sequence (PS-6) 5'-CGUAGC-3' 3'-GC$_{ps}$A$_{ps}$TCG-5'				
A	B	Expt. Tm (°C)	$\Delta\Delta G_{hyb}$ (kcal/mol)	sd (kcal/mol)
	DNA	44	0	
	[Sp, Sp]-DNA	42	-2.60	0.49
DNA	[Rp, Rp]-DNA	40	0.75	0.49
	DNA	43	0	
	[Sp, Sp]-DNA	38	0.28	0.49
RNA	[Rp, Rp]-DNA	42	-1.07	0.49
Sequence (PS-10) 5'-CG$_{ps}$GCCGCCGA-3' 3'-GCCG$_{ps}$GCGGCT-5'				
A	B	Expt. Tm (°C)	$\Delta\Delta G_{hyb}$ (kcal/mol)	sd (kcal/mol)
DNA	DNA	76.55	0	
Sp-DNA	Sp-DNA	77.53	-1.08	0.53
Rp-DNA	Rp-DNA	71.99	1.14	0.56

Table 10: Experimental and predicted hybridization stability for stereodefined phosphorothioate DNA.

T or U also denotes U or T in corresponding RNA or DNA.

The structural parameters were analyzed by using Curves+.[256] Comparison of the parameters for canonical DNA, Rp and Sp indicate that there are no significant differences. NMR experiments also found that the structures of PS modifications are similar to the canonical DNA.[215] This suggest that a small number of PS modifications have little effect on the global helical structures of the duplexes.

The single strands are more sensitive to the modifications. Smith et al. found that ssDNA may prefer an extended conformation, although there is no indication on the detailed structural features;[257] it was hypothesized that ssDNA resembles random coils.[258] The MD simulations allow for a closer look at the single-strand conformations. For PS-6,

DNA, Rp and Sp single strands adopted different global helical conformations, which can be clearly differentiated by the sugar puckering conformations. As shown in Figure 35, Sp adopted C2'-endo conformation while Rp adopted C3'-endo conformation. Additionally, the effect of the Sp and Rp on the conformation is nonlocal, which contrasts with the duplexes. The PS modifications of the first and third linkage changed the conformation of the fifth nucleotide.

Similar trends of the effect of PS modifications were also observed for the 10-mer (Figure 36), while the local sugar puckering conformation does not determine global helical structures. During the simulations, random coil-like structures were observed both different modifications. Some are more extended while others are more condensed.

The change in sugar conformations due to the PS modifications are correlated with the hybridization free energies. DNA duplex and DNA/RNA hybrids typically adopt B-form and A-form conformation, respectively. Sp single strands prefers B-form, so the Sp modifications stabilize the DNA duplex; Rp single strands prefers A-form, so the Rp modifications stabilize the DNA/RNA hybrids. The same conclusion was reached in a previous QM study.[227] While the previous work is more qualitative, our MD simulations produced free energy estimation that can be directly compared to experiments.

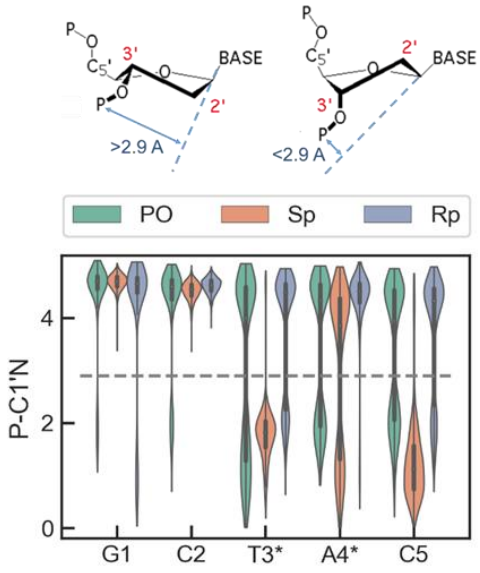


Figure 35: Sugar puckering distance of PS-6 single strand.

“PO” denotes unmodified DNA. “Sp” and “Rp” are PS stereoisomers. “*” denotes the position of the PS on the 5’-end. P-C1’N is the distance between the phosphate P atom and the C1’-N1/N9 line in the previous nucleotide, which is used to quantify sugar puckering conformation. A distance below 2.9 Angstrom indicates B-form while a distance above 2.9 Angstrom indicates A-form.

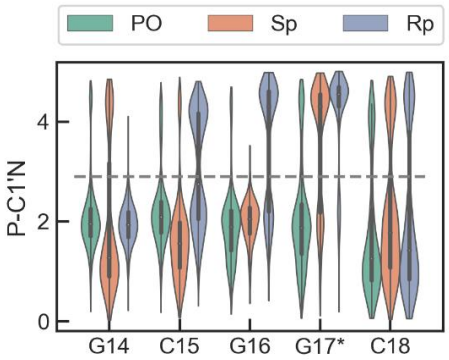


Figure 36: Sugar puckering distance of PS-10 single strand.

See Figure 35 for explanations of the calculation.

3.2 Antagomirs with PS and LNA modifications

Free energy calculations were used to study the effect of PS, LNA and PS-LNA modifications on the antagomirs. The hybridization under consideration is between 8-mer fragment of the miR-22-3p and its antisense. The free energy results and the uncertainties are listed in Table 11. Several combinations of the modifications were considered, which include different isomers and different positions. The uncertainties for PS were larger than expected, because the single strand conformation is very sensitive to the modifications. In the simulation results, PS is mostly unfavorable for the hybridization. The modification with 7 Sp exhibit the larger change in free energy, 4.6 kcal/mol. The modification with 7 Rp has relatively little effect on RNA/DNA hybrids. The different effects of Sp and Rp are also observed in the validation results. The contribution of multiple PS modifications are not simple summations of the contributions of individual modifications. For example, more Rp modifications does not always mean more favorable free energies.

The free energy for LNA is relatively easy to converge. All the LNAs considered contributed to favorable hybridization free energies. The magnitude of the free energy changes correlated with the number of LNAs. The net effect PS-LNA is also favorable for hybridization. Among the isomers of PS-LNA, the R isomer is the more favorable. The addition of PS to LNA generally decreases the stability, although the total effect is still increased stability.

As discussed previously, the single-strand conformation correlate with the hybridization stability. Representative structures for the Sp and Rp modified single strands are shown in Figure 37. Like the PS-6 and PS-10 single strands, Sp and Rp prefers C2’-endo and C3’-endo conformations, which lead to B-form and A-form global helical structures. The sugar puckers measurements can be seen in Figure 38. The single

strands with two isomers do not have well-defined conformations. The canonical single-strand DNA is partially unstacked, which agrees with previous hypothesis on the conformation of single-strand DNAs.[257]

The effect of LNA is related to the restrain on sugar puckering conformations. The single-strand LNAs, including 7 and 4 modifications, shows very stable well-defined global conformations with A-form structure. When 4 LNA modifications are combined with PS modifications, disruption of the global structures was observed.

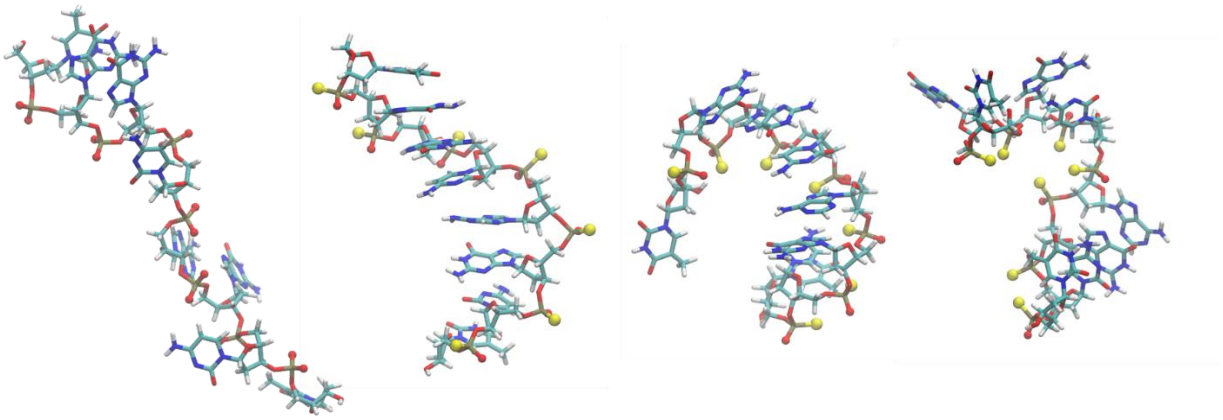


Figure 37: Conformations of single-strand DNA with and without PS modifications.

The structures are DNA (A-0), Sp (A-11), Rp (A-12) and mixed Sp/Rp isomer (A-14), respectively.

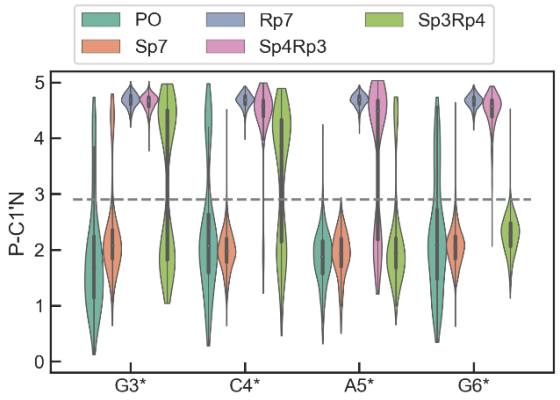


Figure 38: Sugar puckering distance of single-strand antagomirs with PS modifications.

PO stands for phosphate linkage while other antagomirs are described in Table 11.

Sequence	Name	ID	$\Delta\Delta G_{hyb}$ (kcal/mol)	sd (kcal/mol)
5'-T-G-G-C-A-G-C-T-3'	DNA	A-0	0	--
5'-TsGsGsCsAsGsCsT-3'	PS Sp7	A-11	4.6	1.9
5'-TrGrGrCrArGrCrT-3'	PS Rp7	A-12	1.0	2.0
5'-TsGrGsCrAsGrCsT-3'	PS Sp4Rp3	A-13	1.9	1.9
5'-TrGsGrCsArGsCrT-3'	PS Sp3Rp4	A-14	0.0	2.0
5'-T-G-G-C-A-G-C-T-3'	Ln 8	A-20	-4.1	0.7
5'-TsGsG-C-A-GsCsT-3'	Ln8+Sp4	A-21	-4.3	1.0
5'-TrGrG-C-A-GrCrT-3'	Ln8+Rp4	A-22	-3.6	1.1
5'-TsGsGsCsAsGsCsT-3'	Ln8+Sp7	A-31	-4.1	2.0
5'-TrGrGrCrArGrCrT-3'	Ln8+Rp7	A-32	-2.5	2.0
5'-T-G-G-C-A-G-C-T-3'	Ln5	A-40	-2.0	0.5
5'-TsGsG-C-A-GsCsT-3'	Ln5+Sp4	A-41	-3.4	1.0
5'-TrGrG-C-A-GrCrT-3'	Ln5+Rp4	A-42	-2.3	1.0

Table 11: Predicted hybridization stability of miR-22 and PS/LNA antagomirs.

The table is adapted from Ref.[200] with permission. Underscore means LNA; “r” and “s” means Rp and Sp linkage.

3.3 PNA antagomirs

The effect of PNA as a function of sequency length from 8 to 12 was investigated by free energy calculations. For comparison, Mfold hybridization free energies were included.[259,260] Typically, RNA duplexes are more stable than corresponding DNA/RNA hybrids.[261] The results are listed in Table 12. The total free energy change for PNA is more favorable than canonical nucleic acids. However, the predicted changes in individual steps have large variations, while the empirical values are more consistent between the steps. This may be because of the large uncertainties of the free energy calculation as well as small system size. In the simulations, the addition of purines leads

to largest increase in the stability, indicating that PNA further amplifies the strong stacking interactions.

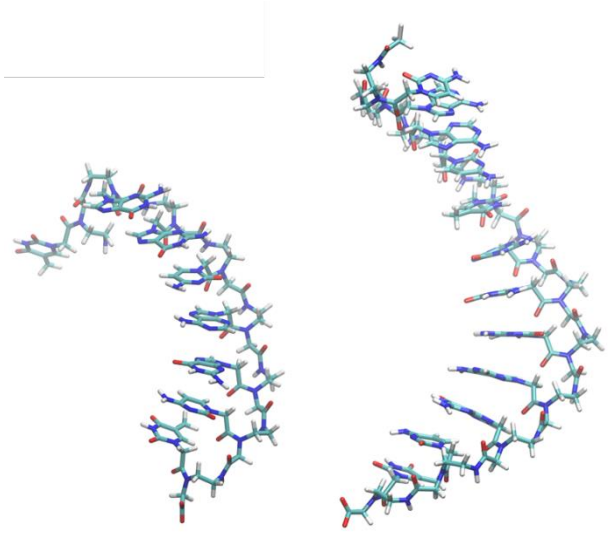


Figure 39: Structures of octamer and dodecamer single-strand PNAs from MD simulations.

The conformation of the PNA single strands is very stable and close to A-form DNA (Figure 39), which may be part of the reason for the enhanced hybridization stability. The stability of single-strand PNAs was also observed in previous short simulations using fixed-charge force field.[220] However, in a recent study employing the early versions of AMBER force field lead to spurious stacking structures,[234] which is also found in simulations of different sequences of RNA tetranucleotides. This structure was considered to be caused by imbalance between hydration and stacking interactions.[241]

n	Sequence	Mfold		Simulation	
		RNA/RNA	DNA/DNA	PNA/RNA	sd
9	5’-CTGGCAGCT-3’	-1.68	-1.97	-0.24	0.18
10	5’-ACTGGCAGCT-3’	-3.56	-3.20	-1.86	0.32
11	5’-AACTGGCAGCT-3’	-5.39	-4.84	-5.98	0.37
12	5’-CAACTGGCAGCT-3’	-6.63	-6.16	-6.90	0.41

Table 12: Predicted and empirical Mfold hybridization free energies (kcal/mol) for PNA-RNA, DNA/RNA hybrids and RNA/RNA duplexes.

The target sequence is 5’-AAGCUGCCAGUUGA-3’. U or T also denotes the corresponding base in DNA or RNA.

4. SUMMARY

The AMOEBA force fields for several modified nucleic acids were developed based on standard protocols. Validation based on experiments of various PS sequences demonstrated the good accuracy of our free energy simulations. For all the sequences studied, PS leads to weaker hybridization while LNA leads to strong stabilization of the hybrids. Combination of LNA and PS also improves the hybridization stability. PNA generally has stronger binding than canonical DNA, while the effect of PNA is different for different nucleotides.

Analysis of the conformations of single strand modified nucleic acids shows that the enhanced or decreased hybridization stability can be largely explained by the conformation preorganization. Preorganization alone cannot explain the relative stability between PS and DNA, since single-strand Rp is closer to the duplex conformation than single-strand DNA but Rp is unfavorable for hybridization. The destabilization effect is likely due to the reduced backbone flexibility of PS. The duplex stability is a result of competition between free energies of single strands and duplexes.

The results also demonstrate that quantitative predictions for the hybridization stability can be achieved by using FEP and polarizable force fields.

Chapter 8: Refinement of AMOEBA RNA force field

1. INTRODUCTION

Molecular dynamics (MD) simulations have been widely used to study the conformational transition of proteins and protein-ligand binding. From the computational point of view, RNAs are more complicated than proteins and the dynamics for RNA conformational transition is longer than that for proteins. Recently, due to the advance in computational hardware, especially general-purpose graphics processing unit (GP-GPU), and improvement in the accuracy of computational models, MD simulations have become increasingly popular for the study of RNAs.

The most popular nucleic acids force field is AMBER. The first versions of AMBER force field, parm94 and parm99, were developed back in the 1990s. These versions were based on quantum mechanical (QM) calculations with relatively low level theories on small model systems.[262] AMBER parm99 performs reasonably well at maintaining the stability of RNAs in a few nanoseconds, which were common simulation time scales at that time.[262] Later, problems of AMBER force field were identified in longer simulations. A study by Varnai and Zakrzewska found serious distortions of DNA duplexes in 50 ns simulations.[262] Several groups have presented reparameterization of the AMBER force fields by utilizing better QM methods, more realistic model compounds, and longer simulations. Pérez et al. derived new α/γ torsion parameters (parmbsc0) to improve the decription of α/γ conformers. The modifications were based on an extended model compound (Figure 40) optimized using MP2/6-31+g(d) and B3LYP/65-31+g(d). Parmbsc0 has better description of DNA duplexes as verified by multiple 200 ns simulations. In addition, it produced stable structures of canonical and noncanonical RNA in short simulations and was able to model a wide range of DNA structures.[262]

Using similar strategies, XXXXX. However, some level of empirical input was required since using better QM methods did not always lead to better force field parameters.

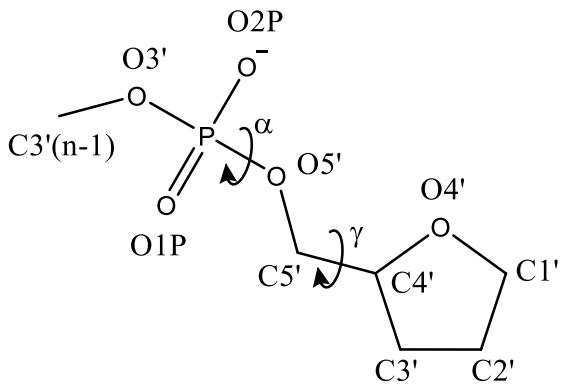


Figure 40: Model compounds for the derivation of α/γ parameters in AMBER parmbsc0.[262]

The figure was prepared using ChemDraw.

AMOEBA nucleic acids force fields were developed in 2018.[56,243,244] The first version of AMOEBA showed good accuracy for a variety of systems, including DNA duplexes in different environment, RNA tetranucleotide, RNA tetraloops, RNA duplexes, and HIV-TAR. However, protein-RNA interfaces and some speficic noncanonical RNA structures were not included the original development. In further testing of selected systems, it was found that AMOEBA does not correctly represent protein-RNA interface and Kink-turn. In this work, the AMOEBA RNA force field was refined to improve the accuracy for these systems.

2. BENCHMARK OF QM METHODS

The quality of QM methods used for force field parameterization has a significant effect on accuracy of force fields. Therefore, the QM methods that has been used for the development of AMOEBA were revisited.

The major QM method used for AMOEBA and many other force fields is MP2,[26,244,262] usually combined with a large basis set or extrapolated to the complete basis set (CBS) limit. MP2 is an ab initio method with good accuracy for a variety of systems. However, MP2 is known to be problematic for dispersion-dominated interactions.[103,263] Several empirical methods based on MP2 have been proposed to improve the accuracy for dispersion, including spin-component-scaled (SCS) MP2[263] and dispersion-corrected MP2.[264,265] Another issue with MP2 calculation is its poor scaling with system size. MP2 calculation with large basis set is limited to relatively small systems. An althernative class of methods is density functional methods (DFT), which typically has better scaling and reasonable accuracy.

Recent development of DFT methods, including empirical dispersion corrections, has significantly improved the accuracy of DFT. In many cases, DFT has shown better accuracy than MP2.[152] DFT methods are usually classified into five rungs of Jacob's ladder, with increasing complexity, computational cost and accuracy. Methods in the fifth and highest ladder includes correlation calculated by second order perturbation or other methods and are therefore not cheaper than MP2. Additional, recent development of more efficient approximation to MP2, such as local MP2 and density fitting MP2, has significantly reduced the computational cost of MP2. As a result, MP2 can be faster than some DFT methods for medium sized molecules (10-100 atoms). Therefore, the question is to choose an efficient and accuracy method from all DFT or MP2-based methods.

A range of QM methods were benchmarked on common data set for their accuracy and computational speed (Figure 60, Figure 61 and Figure 62). The two double-hybrid DFT methods with empirical dispersion correction, DSD-BLYP-D3BJ and DSD-PBEP86-D3BJ have very good accuracy and a computational cost similar to that of MP2. The highly parameterized ωB87 family of DFTs from Head-Gordon[103] also has good accuracy, but its implementation in Psi4 is less efficient due to the calculation of range-separated HF exchange. Traditional hybrid DFTs (such as B3LYP, PBE) combined with minimal basis set are a good alternative for very large systems because of their low computational cost.

3. OPTIMIZATION OF VDW PARAMETERS

3.1 Nucleobase vdW parameters

Model compounds with equilibrium geometries was taken from the literature[244] and re-optimized by MP2.

Model compounds with non-ideal geometry were generated with the following procedure:

1) Extraction of protein-NA interface from crystal structures.

   5334 PDBs containing both protein and NA were downloaded from RCSB PDB. The residues at protein-NA interface were extracted.

   The fragments were converted to dimers of model compounds.

2) Clustering and cleaning.

   Starting from 68 dimers, 113251 structures, a hierarchical clustering was used to select representative structures. A RMSD cutoff of 2.5 Å was used. The RMSD was calculated on the dimer after aligned on the first monomer.

Spurious bond lengths with E_bond > 10 kcal/mol were relaxed by restrained optimization using AMOEBA.

Short intermolecular distances (<1.8 Å) were relaxed by rigid distance scan with DFT-D/PCM.

The hydrogen positions were optimizied by DFT-D/PCM or PM6/PCM.

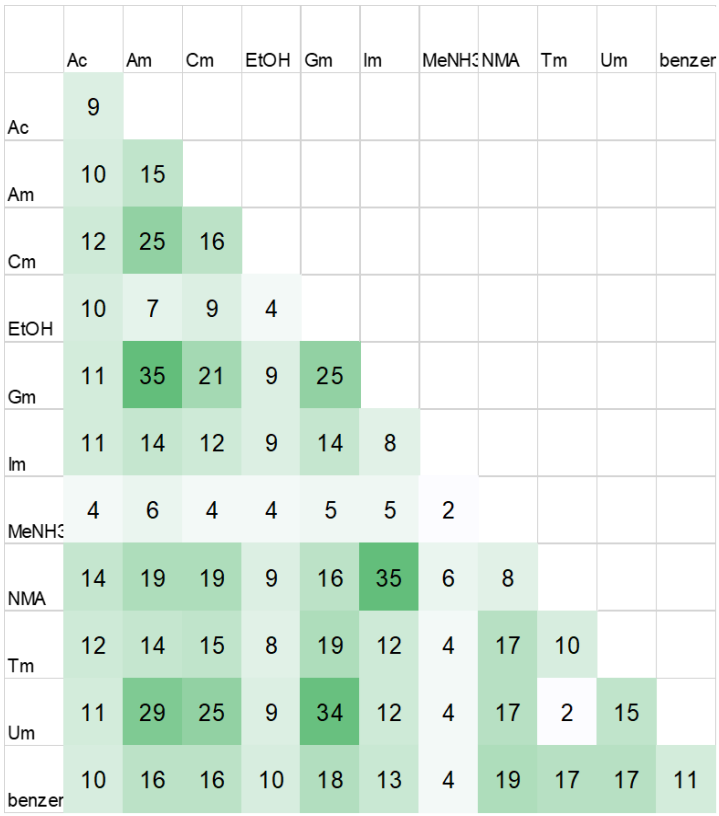
3) Interaction energy calculation.

   Interaction energies were calculated by DSD-BLYP-D3BJ/def2-QZVP from all 68 dimers and 862 structures from the last step.

   High energy structures (Eint > 10 kcal/mol) were discarded.

4) Final data set contains 66 dimers, 862 structures.

   Summary of the data set for the non-ideal geometry is shown in Figure 41.



	Ac	Am	Cm	EtOH	Gm	Im	MeNH3	NMA	Tm	Um	benzer
Ac	9										
Am	10	15									
Cm	12	25	16								
EtOH	10	7	9	4							
Gm	11	35	21	9	25						
Im	11	14	12	9	14	8					
MeNH3	4	6	4	4	5	5	2				
NMA	14	19	19	9	16	35	6	8			
Tm	12	14	15	8	19	12	4	17	10		
Um	11	29	25	9	34	12	4	17	2	15	
benzer	10	16	16	10	18	13	4	19	17	17	11

Figure 41: Summary of dimer structures in the protein-NA interface data set.

Am, Tm, Gm, Um, Cm denote methylated A, T, G, U, C nucleobases, respectively.

First, several QM methods along with amoebabio18 parameters were compared on the interaction energy of selected base pairs (Figure 42 and Figure 43). For the hydrogen bonding interaction between A-G base pair, the considered QM methods except SCS-MP2 have similar interaction energies. SCS(MI)-MP2 was designed to have better dispersion interaction at the cost of slightly worse accuracy for other types of interactions. The amobeabio18 parameter is significantly more repulsive than most of the QM methods. The interaction energy between A-G base pair as a function of the y-displacement is also investigated (Figure 43). All the QM methods have similar slope,

even though SCS(MI)-MP2 has weaker interaction energies. The ameobabio18 parameter has relatively flat interaction energy curve compared to QM, which means that the interaction is less specific.

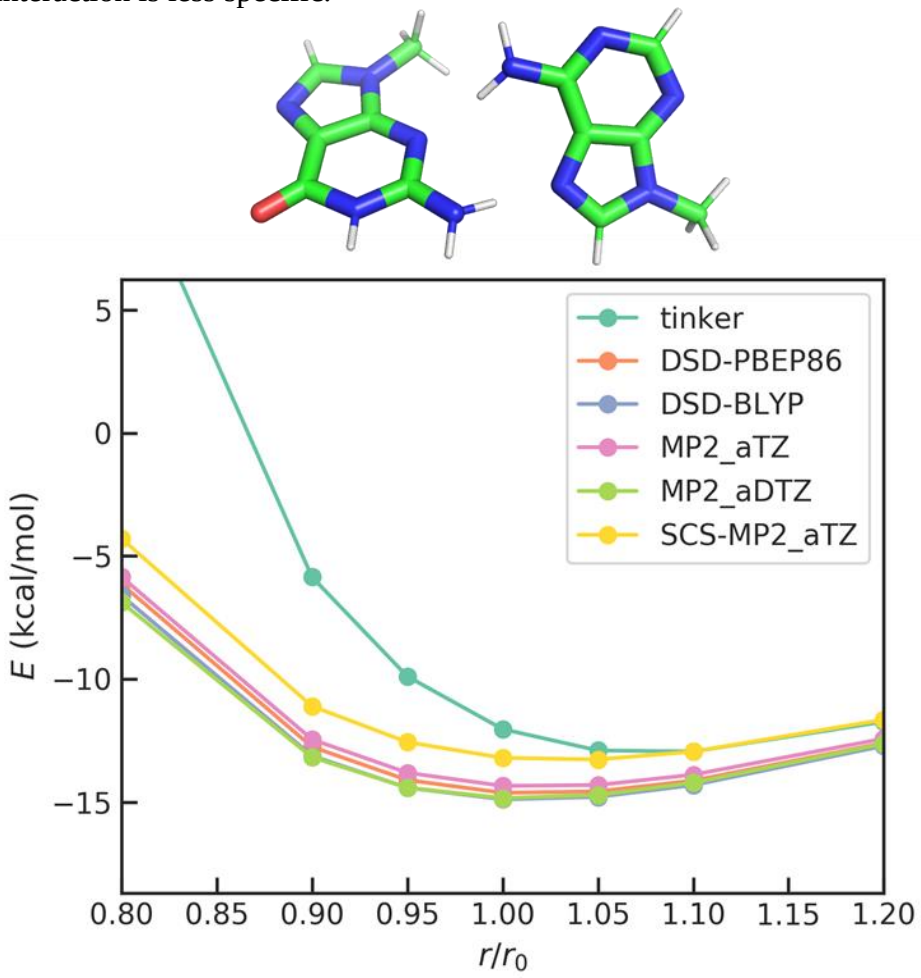


Figure 42: Comparison of force field and QM methods for Adenine-Guanine cis-Hoogsteen-Sugar edge base pair as a function of the relative distance.

"tinker" means AMOEBA force field with the amoebabio18 parameter calculated by using Tinker; "DSD-PBEP86" means DSD-PBEP86-D3BJ/def2-QZVP; "DSD-BLYP" means DSD-BLYP-D3BJ/def2-QZVP; "MP2_aTZ" means MP2/aug-cc-pVTZ; "MP2_aDTZ" means MP2/CBS extrapolated from aug-cc-pVDZ and aug-cc-pVTZ. "SCS-MP2_aTZ" means SCS-MP2/aug-cc-pVTZ; The distance "r_0" is the equilibrium distance. QM interaction energy was calculated by using Psi4 with counterpoise correction.

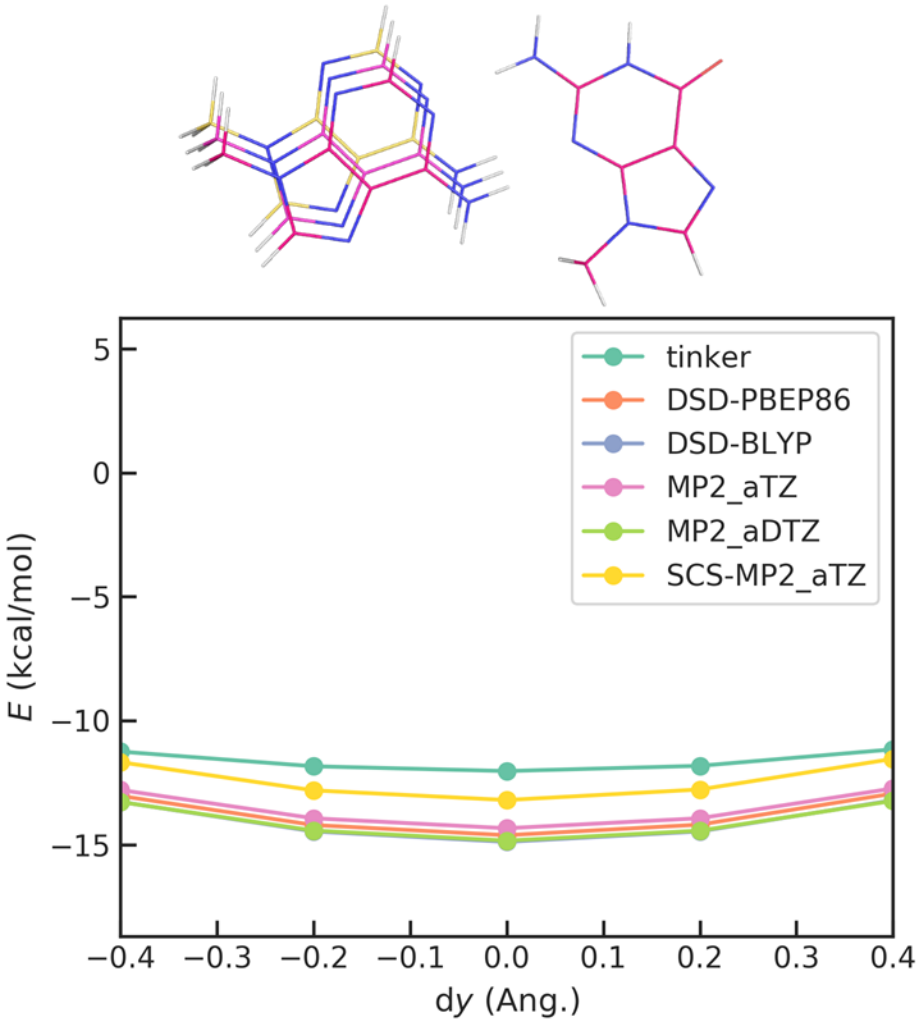


Figure 43: Comparison of force field and QM methods for Adenine-Guanine cis-Hoogsteen-Sugar edge base pair as a function of the shift along y-axis.

See Figure 42 for nomenclature and computational details.

DSD-BLYP-D3BJ was chosen as the reference DFT method for other systems due to its accuracy in our benchmark and previous studies.[152] The error of amoebabio18 parameter for all the base pair interactions is shown in Figure 44. While the interaction for many base pairs can be predicted within an error of 2 kcal/mol, the base-pairs

containing G have significantly larger errors than the others in terms of both mean error (ME) and root mean squared error (RMSE). This can be more clearly seen in the distribution of errors for each base pair (Figure 45). The average error for A-G, C-G, G-G, G-T, G-U ranges from 2 kcal/mol to 5 kcal/mol, while the average errors for other base pairs are around 1 kcal/mol.

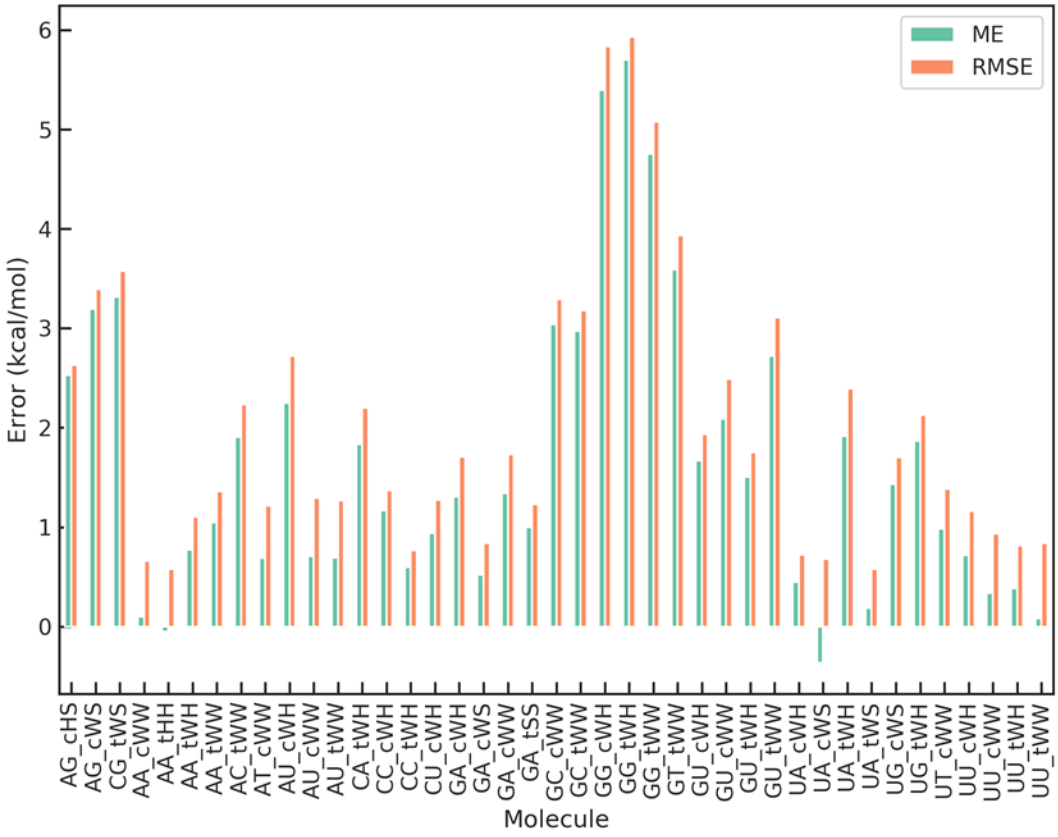


Figure 44: Error of AMOEBA force field for base-pair interaction evaluated by the QM data (DSD-BLYP-D3BJ) compiled in this work.

“ME” and “RMSE” denote mean error and root mean squared error, respectively.

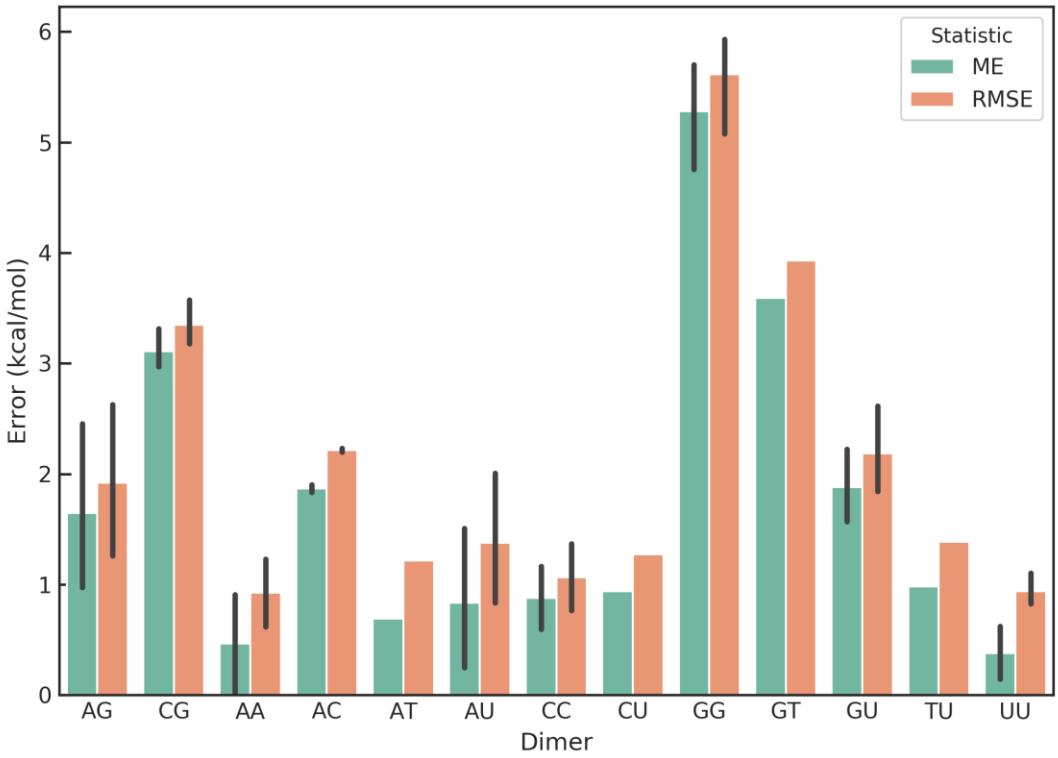


Figure 45: Distribution of errors of AMOEBA force field for base-pair interaction for each base pair.

"ME" and "RMSE" denote mean error and root mean squared error, respectively.

The AMOEBA vdW parameters of the nucleobases were optimized to reduce the errors for these interactions. Three sets of optimization schemes are considered. In the first two schemes, the number of atom types is the same as the amoebabio18 parameter, and different degrees of regularization was used to keep the optimized parameter close to the original parameters. In the third scheme, separate atom types were used to describe Guanine. The optimized parameters for the first and the third schemes are listed in Table 13. The changes in parameters are less than 5% from amoebabio18. The vdW radii are smaller in all optimized parameters than in amoebabio18, since the base-pair interactions

in amoebabio18 are too weak. The fitting results are shown in Figure 46 and Figure 47. The RMSEs can be reduced by half for the base pairs containing Guanine. As expected, using weaker regularization or more atom types can further reduce the fitting error. However, this is a risk of overfitting. Therefore, the three sets of optimized parameters need to be tested by other data., Using separate atom types for Guanine significantly improves the mean errors for many base pairs (Figure 47), while the mean errors are roughly the same when different regularizations are used. This demonstrates the benefits of including more parameters.

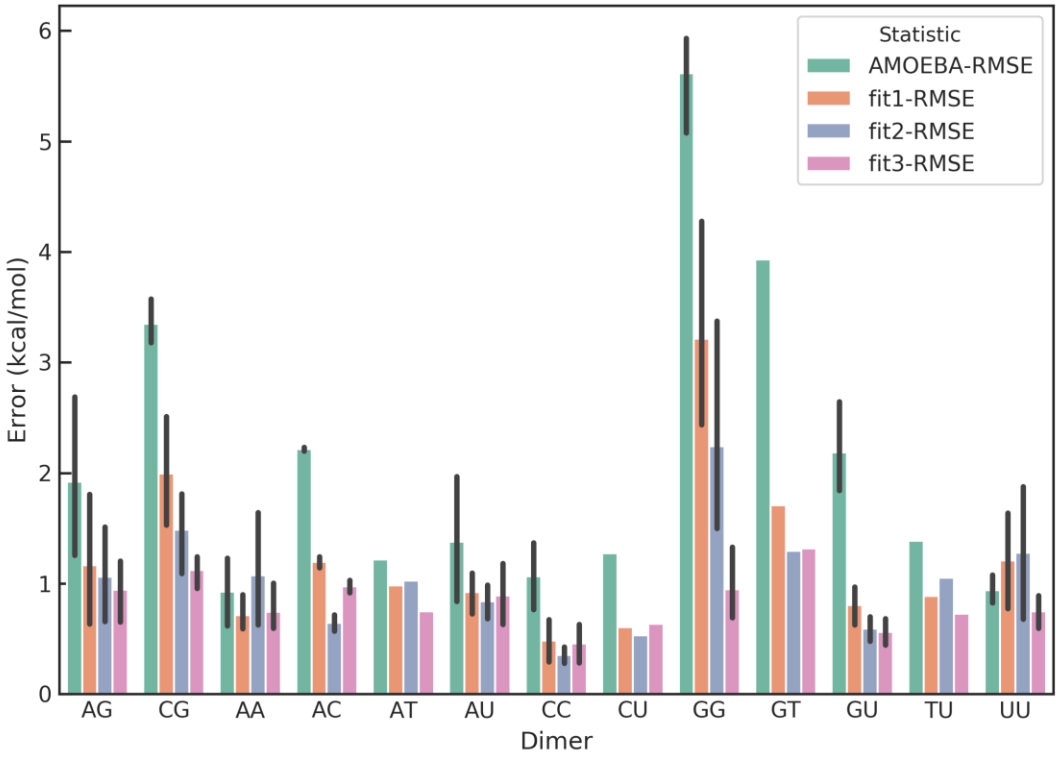


Figure 46: RMSE of AMOEBA force field for base-pair interactions with the amoebabio18 parameter and three sets of new parameters.

“fit1” was fitted using strong regularization; “fit2” was fitted using weak regularization; “fit3” was fitted using separate atom types for Guanine in addition to weak regularization.

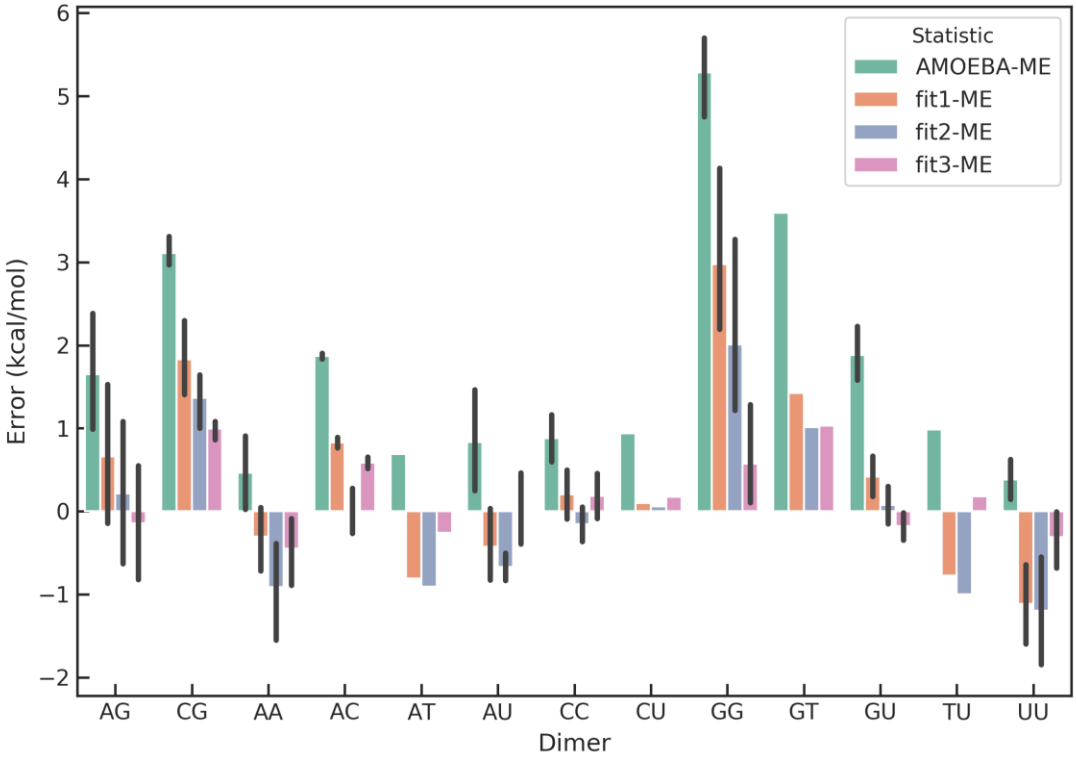


Figure 47: Mean error of AMOEBA force field for base-pair interactions with the amoebabio18 parameter and three sets of new parameters.

See Figure 46 caption for explanation of different sets of parameters.

Vdw type	Atom	amoebabio18	fit1	fit3[1]
54	U/T N3, G/C N1	3.700 0.127	3.600 0.124	3.700 0.127 3.560 0.127
50	A N3/N1, G/C N3	3.640 0.127	3.530 0.124	3.560 0.120 3.500 0.125
49	A/G N7	3.640 0.127	3.530 0.124	3.560 0.120 3.500 0.125
58	C O2, G O6, U/T O2/O4	3.350 0.129	3.250 0.124	3.250 0.124 3.200 0.120
53	A H6, C H4, G H2	2.650 0.020	2.620 0.018	2.650 0.020

Table 13: Nucleobase vdW sigma (Å)/epsilon (kcal/mol) in amoebabio18 and optimized parameters.

Model	CV[2]	Training		Test[3]	
		Stack[4]	HBond[5]	Base[6]	Base-AA[7]
AMOEBA		1.466	3.775	1.895	1.309
fit1	1.077	0.86	1.317	0.829	1.193
fit3	0.851	0.887	1.033	0.856	1.194

Table 14: Evaluation of AMOEBA parameters on cross-validation, training, and test sets.

The errors are RMSE in kcal/mol.

[1] "fit3" used separate atom types for A/U/C and G. The first line in each cell is for A/U/C, and the second line is for G.

[2] K-fold cross-validation with k=5.

[3] The test set consists of non-ideal geometries

[4] "Stack" means base-base stacking

[5] "HBond" means base pairing

[6] "Base" includes both "Stack" and "Hbond";

[7] Base-AA means base-amino acid interactions

The performance of the optimized parameters is evaluated based on cross-validation (CV) and test data sets (Table 14). “fit3” has significantly better training error for H-bonding interactions and better CV error than “fit1”. However, “fit3” has slightly worse training error for base-base stacking interactions and test errors. According to the performance on the test sets, “fit3” was chosen to be the final optimized parameters.

3.2 Nucleobase-protein vdW parameters

The hydrogen bonds between Asp/Glu side chains and nucleobases are a common motif in protein-nucleic acid interfaces. It was found that even with the optimized nucleobase vdW parameters, some of these hydrogen bonds were not stable during long MD simulations. Therefore, the accuracy of AMOEBA for the protein-NA test set was further examined (Figure 48). Consistent with the MD simulations, the AMOBEA interaction energies for the carboxylate groups are significantly smaller than the QM interactions. This may be attributed to the partial covalent bonding nature of these interactions. Special vdW pair parameters between carboxylate O and NH were included to improve the accuracy for these interactions.

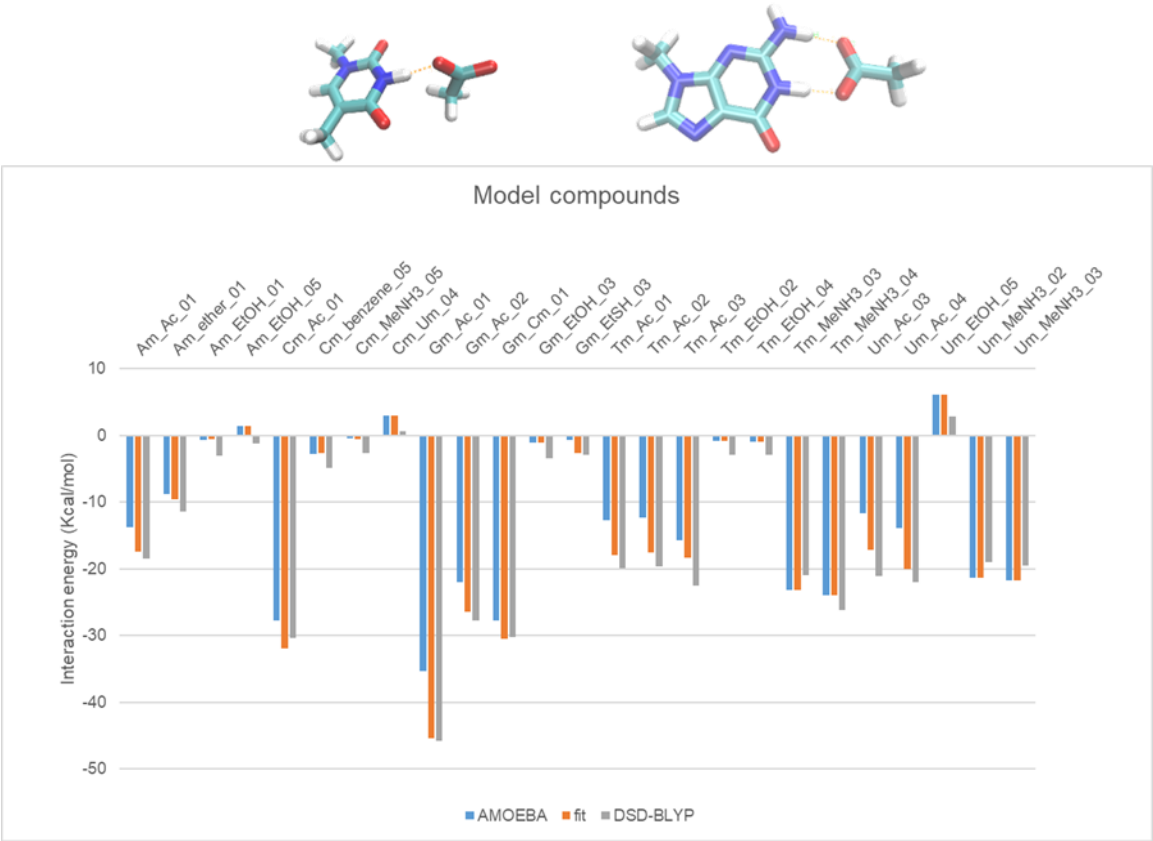


Figure 48: Comparison of AMOEBA and QM on base-amino acid interactions.

"fit" contains the "fit1" parameters in Table 14 and special vdW pair parameters between carboxylate O in Asp/Glu and NH in nucleobases.

4. Optimization of Asn torsion parameters

During long MD simulations of the U1A protein-RNA interface, it was also found that the rotamer conformation of some Asn residues in the interface was different from those in the crystal structures, which leads to the breaking of some hydrogen bonds. Therefore, the torsion parameters of Asn were optimized based on conformations from torsion scan and from the rotamer library and ωB97xd/6-311++g(2d,2p) level of theory. The original and the optimized torsion parameters were compared on the 1-D torsion

profile of χ1 (Figure 49) and on the conformations sampled from MD simulations using amoeabio18 (Figure 50). For the 1-D torsion profile, noticeable improvements are observed for both alpha-helix or beta-sheet conformations, although the profiles for both conformations cannot be exactly reproduced at the same time.

For the conformations sampled from MD simulations, the original parameters have significant errors in the relative energies of the two major conformations. The energies for both conformations are predicted to be similar by the original parameters, while QM calculations indicate that one conformation is on average 4 kcal/mol lower in energy than the other conformation. This error explains the observation of a different conformation from previous MD simulations. The optimized parameters have much better agreement with QM on these conformations, although the accuracy is slightly worse than the parameters directly fitted using the QM data. Considering that the optimized parameters are obtained using a general procedure (1-D torsion scan and rotamer library), they are considered both robust and accurate for simulations of different systems.

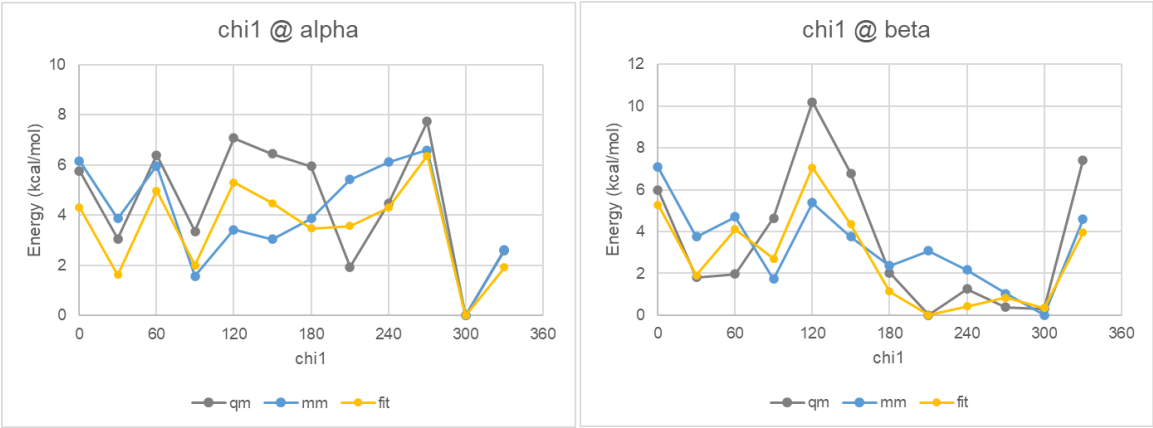


Figure 49: Results of Asn side chain torsion fitting.

The torsion parameters fitted using structures from rotamer library and torsion scan. There are 11 rotamers of Asn. For each rotamer, a 3 by 3 scan of chi1/chi2 with spacing of 30 deg. was used to generate a total of 99 structures from rotamer library. In addition, 1-D torsion scan for chi1 or chi2 with spacing of 30 deg., starting from alpha-helix or beta-sheet conformation was performed. “mm” means amoebabio18 torsion parameter, and “fit” means the optimized parameter.

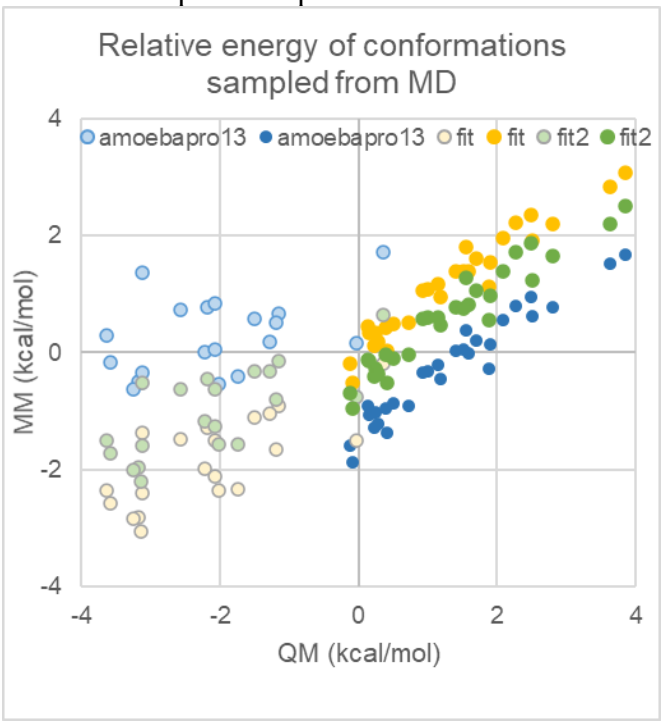


Figure 50: Comparison of relative energies of Asn with conformations sampled from MD simulations.

“fit” represents the torsion parameter directly fitted to reproduce the relative energies of these conformations; “fit2” represents the torsion parameters fitted using structures from rotamer library and torsion scan.

5. Simulations of protein-RNA interface.

The final AMOEBA parameters with modifications on nucleobase vdW parameters, special vdW pair parameters between NH and carboxylate and Asn sidechain torsions were used for simulations of two problematic RNA interfaces, U1A and FBF.[240] The U1A protein is bound to an RNA hairpin. It is part of the U1 snRNP spliceosomal

cmplex.[240,266] The crystal structure of U1A contains a 21-nucleotide RNA sequence. Previous simulations with two fixed-charge force fields could not stabilize the protein-RNA interface. An altered hydrogen-bond network was formed and stabilized during the simulations. The *C. elegans Pumilio* FBF-2 protein is bound to *gld-1* FBEa RNA. The protein is responsible for the regulation of messenger RNA.[266] The FBF interface is composed of interaction between nucleobases and protein, while the RNA backbone is exposed to the solvent.[266] Previous simulations with fixed-charge force fields could not stabilize the interface. The hydrogen bonds in the interface were gradually broken in those simulations, which is mainly caused by the conformational changes of RNA.[266] A recently modification of AMBER force field that includes generalized H-bond function led to improved description of the U1A interface,[267] while no further work on FBF has been done.

The results of the original AMOEBA parameters and optimized parameters are shown in Figure 51 and Figure 52. In FBF simulations, the U1-Asn500 and G2-Glu457 hydrogen bond are more stable with the revised parameters (Figure 51). The distances of hydrogen bonds involving Lys still have wide distributions, which can be attributed to the flexibility of the Lys side chains. For U1A, the original parameters lead to the breaking of the interface in 10 ns. The interface is stable in multiple 60-ns simulations (Figure 52).

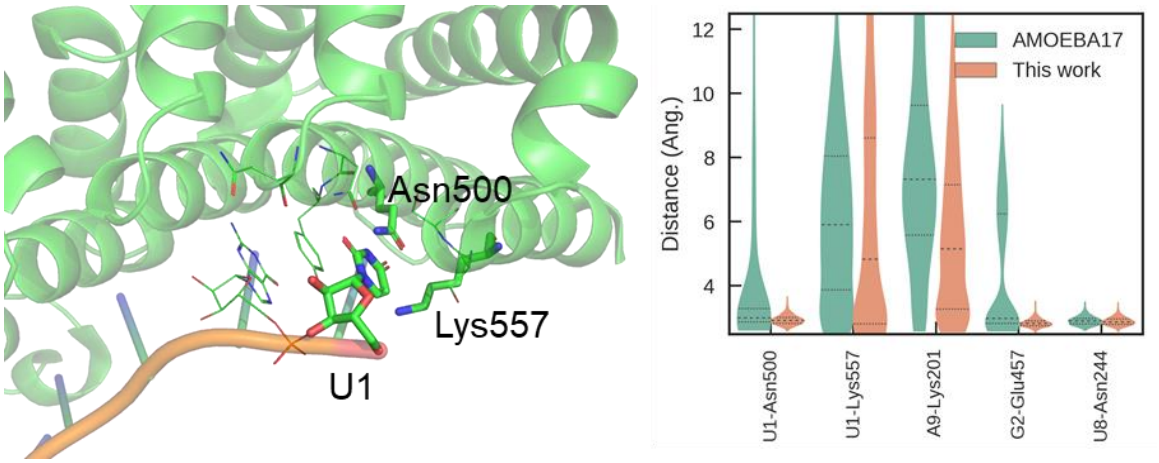


Figure 51: Hydrogen-bond distances from MD simulations of FBF.

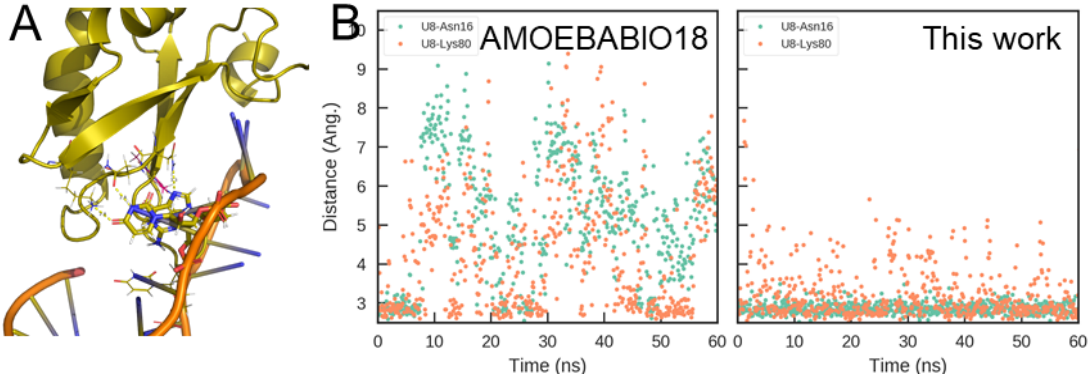


Figure 52: Hydrogen-bond distances from MD simulations of U1A.

(A) structure of the protein-RNA interface. (B) time evolution of the hydrogen-bond distances in MD simulations with original and optimized parameters.

6. SUMMARY

The AMOEBA RNA force field was refined to improve the accuracy for modeling protein-RNA interfaces. The most accurate DFT methods were chosen as a reference for the intermolecular between nucleobases, water and small organic molecules. The nucleobase parameters were optimized by fitting to a large QM dataset including

both optimized and non-ideal geometries and validated by cross-validation and separate test sets. Special vdW pair parameters were included to better represent hydrogen-bond interactions between nucleobases and carboxylate groups. The Asn torsion parameters were optimized based on conformations from torsional scan and rotamer library. The optimized AMOEBA parameters achieve much better accuracy for the problematic protein-RNA interfaces identified in previous work. Considering the procedure used in the force field optimization, this version of AMOEBA force field is considered a general improvement for simulations of both RNAs and protein-RNA interfaces.

Chapter 9: Conclusion

Metal ions and nucleic acids are essential components of biology. The molecular mechanisms of many of their biological functions remain unclear. From a computational perspective, metal ions and nucleic acids are challenging for both classical models and quantum mechanical (QM) methods. Classical models can be problematic for the interaction of ions. QM is very useful for small and simplified systems, while metal ions and nucleic acids often reside in complex and dynamic environments.

This dissertation is concerned with the application of polarizable force fields to study the thermodynamics of ion and nucleic acid recognition. In Chapter 3, protein-ion interactions were analyzed by QM methods and the importance of many-body energy and charge transfer for ion recognition was revealed. In Chapters 2 and 8, accurate models for metal ions and nucleic acids were developed. In Chapters 4 and 5, successful prediction of ion binding thermodynamics was achieved by using polarizable force fields. In Chapter 7, the effect of chemical modifications on DNA/RNA hybridization stability was explained by free energy calculation with polarizable force fields.

Most of these studies started out with the aim of making reliable predictions by incorporating the polarization effect. Somewhat surprisingly, polarization proved to completely change our perception of some important phenomena. The preference of proteins to bind Ca^{2+} instead of Mg^{2+} was puzzling because Mg^{2+} has stronger electrostatic interactions and was considered a better binder. However, Mg^{2+} also has stronger, unfavorable polarization interaction with the protein, which was overlooked. Simulations demonstrated that polarization plays a decisive role in the recognition of Ca^{2+} by proteins. The ion conduction mechanism in potassium channels was another open question because the direct contacts between four K^{+} ions in the crystal structures could

not be reconciled with existing theoretical models. It was hypothesized under physiological conditions, some of the K^+ binding sites are vacant or occupied by water molecules. These hypotheses did not account for polarization effect, which would significantly reduce the repulsion between K^+ ions and facilitate ion conduction at ion-saturated states. Future applications of polarizable force fields, or perhaps even more efficient and accurate models, will unquestionably lead to more exciting discoveries on the microscopic mechanisms of biomolecular recognition.

Appendices

1. Optimized AMOEBA+ ion parameters

```
atom       6    6    Li+   "Sodium Ion Na+"        3     6.940    0
atom       7    7    Na+   "Sodium Ion Na+"       11    22.990    0
atom       8    8    K+    "Potassium Ion K+"     19    39.098    0
atom       9    9    Rb+   "Rubidium Ion Rb+"     37    85.468    0
atom      10   10    Cs+   "Rubidium Ion Rb+"     55    85.468    0
atom      11   11    Mg+   "Magnesium Ion Mg+2"   12    24.305    0
atom      12   12    Ca+   "Calcium Ion Ca+2"     20    40.078    0
atom      13   13    Zn+   "Zinc Ion Zn+2"        30    65.390    0
atom      14   14    F-    "Fluoride Ion F-"       9    18.998    0
atom      15   15    Cl-   "Chloride Ion Cl-"     17    35.453    0
atom      16   16    Br-   "Bromide Ion Br-"      35    79.904    0
atom      17   17    I-    "Iodide Ion I-"        53   126.904    0
vdw 6 1.97547           0.10762
polarize     6          0.028          0.390  0.25000
ct           6      0.33296    8.00000
cp           6        53.27496  3.0
vdw 7 2.25062           1.34255
polarize     7          0.080 0.050   0.25
ct           7      2.44324    2.39986
cp           7         7.00666  11.0
vdw 8 3.20660           1.05209
polarize     8          0.780 0.050   0.50
ct           8      8.86466    2.87719
cp           8         7.66944  19.0
vdw 9 3.57851           0.99996
polarize     9          1.350          0.390  0.62000
ct           9      4.27606    3.00000
cp           9         5.82760  37.0
vdw 10 4.00452           0.92132
polarize     10         2.260          0.390  0.70000
ct           10    14.37960    3.20000
cp           10        5.18877  55.0
vdw 14 4.45841           0.01156
polarize     14         1.350          0.390  0.70000
ct           14    12.49591    6.57160
cp           14        9.17235  9.0
vdw 15 4.72504           0.18121
polarize     15         4.000 0.390   0.550
ct           15     2.60000    2.50000
cp           15        3.040   17.0
vdw 16 5.02577           0.19154
polarize     16         5.650          0.390  0.70000
ct           16    10.88850    3.00782
cp           16        2.35266  9.0
vdw 17 5.26907           0.27416
polarize     17         7.250          0.390  0.70000
ct           17     7.57976    2.47148
cp           17        2.35059  11.6
```

Table 15: Optimized AMOEBA+ ion parameters in Tinker format.

2. SUMMARY OF MODIFICATIONS TO THE AMOEBA FORCE FIELD PARAMETERS

2.1 Modifications included in the amoebabio18 parameters

```
atom         10    4    HN    "Alanine HN"                 1     1.008    1
atom         11    5    O     "Alanine O"                  8    15.999    1
atom          8    7    CA    "Alanine CA"                 6    12.011    4
atom        137    8    C     "Aspartate CB"               6    12.011    4
atom        139   30    C     "Aspartate CG"               6    12.011    3
atom        140   31    O     "Aspartate OD"               8    15.999    1
atom        153    8    C     "Glutamate CB"               6    12.011    4
atom        157   30    C     "Glutamate CD"               6    12.011    3
atom        158   31    O     "Glutamate OE"               8    15.999    1
atom        357   98    Mg+   "Magnesium Ion Mg+2"        12    24.305    0
atom        358   99    Ca+   "Calcium Ion Ca+2"          20    40.078    0
atom        359  100    Zn+   "Zinc Ion Zn+2"             30    65.380    0

multipole   139  140 -140            1.01811
                                   -0.00488    0.00000   -0.15412
                                    0.00000
                                    0.00000    0.00000
                                    0.00000    0.00000    0.00000
multipole   140  139  137             -0.85879
                                   -0.08949    0.00000   -0.07764
                                    0.00000
                                    0.00000    0.00000
                                    0.00000    0.00000    0.00000

polarize    139         1.3340    0.3900    137  140
polarize    140         1.2000    0.3900    139
polarize    157         1.3340    0.3900    155  158
polarize    158         1.2000    0.3900    157
polarize    357         0.0800    0.1150
polarize    358         0.5500    0.1800
vdw         30             3.8200    0.1060
vdw         31             3.5500    0.0950
vdw         98             2.9000    0.2800
vdw         99             3.5900    0.3500
vdwpr        4   31         3.1000    0.0400
vdwpr        5   98         3.0000    0.1530
vdwpr        5   99         3.2700    0.1750
vdwpr        5  100         2.8900    0.1750
torsion      1    7    8   30     -2.900 0.0 1   1.800 180.0 2   0.000 0.0 3
torsion      3    7    8   30     -6.950 0.0 1  -1.150 180.0 2   0.000 0.0 3
torsion      7    8   30   31      0.000 0.0 1   1.700 180.0 2   0.000 0.0 3
torsion      1    7    8    8     -2.280 0.0 1   0.970 180.0 2   3.700 0.0 3
torsion      3    7    8    8      0.160 0.0 1   1.655 180.0 2  -2.520 0.0 3
torsion      8    8   30   31      0.000 0.0 1   1.460 180.0 2   0.000 0.0 3
```

Table 16. Modified AMOEBA parameters for Mg^{2+}, Ca^{2+}, Zn^{2+}, Asp and Glu.

Mg2+ and Ca2+ parameters were optimized by using MP2/CBS ion-water dimer data and experimental hydration data. Carboxylate polarizability and quadrupole parameters were adjusted based on acetate-water and acetate-Mg^{2+}/Ca^{2+} dimer interactions and acetate hydration data. Mg^{2+}/Ca^{2+} carbonyl vdW special pair parameters were based on dimer interaction. The above changes are described in Ref. [73]. Zn^{2+} carbonyl vdW special pair parameters were based on a similar protocol using unpublished data.

Acetate-amide vdW special pair parameters were based on MP2/CBS dimer interaction energy. Asp and Glu side chain torsion parameters were based on energy scan using ωB97x-D/6-311++G(2d,2p)/PCM. The above changes have been described in Ref. [106].

2.2 Modifications not included in the amoebabio18 parameters (April 2021)

```
atom          7    1    N     "Alanine N"                 7    14.007    3
atom          9    3    C     "Alanine C"                 6    12.011    3
atom         12    6    H     "Alanine HA"                1     1.008    1
atom        140   31    O     "Aspartate OD"              8    15.999    1
atom        147    8    C     "Asparagine CB"             6    12.011    4
atom        149    3    C     "Asparagine CG"             6    12.011    3
atom        158   31    O     "Glutamate OE"              8    15.999    1
atom        190   37    N     "Lysine NZ"                 7    14.007    4
atom        191   38    HN    "Lysine HN"                 1     1.008    1
atom        200   37    N     "Lysine (Neutral) NZ"         7    14.007    3
atom        201   38    HN    "Lysine (Neutral) HN"         1     1.008    1
atom        219   37    N     "Ornithine NE"              7    14.007    4
atom        220   38    HN    "Ornithine HE"              1     1.008    1
atom        234   31    O     "C-Terminal COO-"           8    15.999    1
atom        253   47    N     "Adenine N6"                7    14.007    3
atom        255   49    N     "Adenine N7"                7    14.007    2
atom        256   50    N     "Adenine N3"                7    14.007    2
atom        257   50    N     "Adenine N1"                7    14.007    2
atom        260   53    H     "Adenine H61"               1     1.008    1
atom        261   54    N     "Cytosine N1 RNA"           7    14.007    3
atom        262   54    N     "Cytosine N1 DNA"           7    14.007    3
atom        264   47    N     "Cytosine N4"               7    14.007    3
atom        268   50    N     "Cytosine N3"               7    14.007    2
atom        270   58    O     "Cytosine O2"               8    15.999    1
atom        272   53    H     "Cytosine H41"              1     1.008    1
atom        277   54    N     "Guanine N1"                7    14.007    3
atom        278   47    N     "Guanine N2"                7    14.007    3
atom        282   49    N     "Guanine N7"                7    14.007    2
atom        283   50    N     "Guanine N3"                7    14.007    2
atom        284   61    H     "Guanine H1"                1     1.008    1
atom        285   58    O     "Guanine O6"                8    15.999    1
atom        287   53    H     "Guanine H21"               1     1.008    1
atom        293   54    N     "Thymine N3"                7    14.007    3
atom        296   58    O     "Thymine O2"                8    15.999    1
atom        298   58    O     "Thymine O4"                8    15.999    1
atom        299   61    H     "Thymine H3"                1     1.008    1
atom        303   54    N     "Uracil N3"                 7    14.007    3
atom        306   61    H     "Uracil H3"                 1     1.008    1
atom        308   58    O     "Uracil O2"                 8    15.999    1
atom        309   58    O     "Uracil O4"                 8    15.999    1
vdw          54              3.6000    0.1240
vdw          50              3.5300    0.1240
vdw          49              3.5300    0.1240
vdw          58              3.2500    0.1240
vdw          53              2.6200    0.0180     0.880
vdwpr  31    54              3.4400    0.0930
vdwpr  31    47              3.4400    0.0930
vdwpr  31    61              3.0700    0.0320
vdwpr  31    53              3.0700    0.0320
vdwpr  37    58              3.6800    0.1300
vdwpr  38    58              2.9700    0.0270
torsion       1    7    8    3   -0.74531 0.0 1 -0.39366 180.0 2 -0.22780 0.0 3
torsion       3    7    8    3    0.91766 0.0 1 -0.17805 180.0 2 0.42581 0.0 3
torsion       6    7    8    3   -0.28061 0.0 1   0.000 180.0 2   0.000 0.0 3
```

Table 17. Modified AMOEBA parameters for nucleobases, Asn, Asp, Glu and Lys.

Base vdW parameters were optimized by using DFT data on a large dataset containing base stacking, base pairing, base-water, and base-amino acids dimers. Base-carboxylate and lysine-O vdW special pairs were also added by using DFT ωB97x-D/6-311++G(2d,2p) dimer interaction data. Asn sidechain torsion parameters were based on DFT calculation of conformations from torsion scan and rotamer library. These modifications are described in Chapter 8.

3. FREE ENERGY AND ENTHALPY DATA FOR CALCIUM AND MAGNESIUM BINDING

	Nr. of first-shell D/E	Expt. $\Delta\Delta G$	$\Delta\Delta G$	$\Delta\Delta H$	$\Delta\Delta H_{\mathrm{pol}}$	$\Delta\Delta H_{\mathrm{frz}}$	$-T\Delta\Delta S$	$\sigma(\Delta\Delta H)$
5CPV	4	-5.6	-8.3	-10.2	-18.8	8.6	1.9	2.1
1B8L	4	-1.6	-3.1	8.9	-13.9	22.8	-12.0	2.1
4ICB	3	-6.2	-7.2	-20.1	-13.7	-6.3	12.9	2.5
4IHB	2	-1.7	-3.4	-5.2	-17.9	12.7	5.5	3.5
1ZOO	1	1.7	0.2	-18.1	-24.5	6.5	14.7	3.4
Acetate		-0.1	0.2	-1.9	-14.2	12.3	2.1	2.4

Table 18: Enthalpy and entropy contributions to relative binding free energy (kcal/mol).

Reactant		Product		$\Delta\Delta G_{BIND}$ (kcal/mol)					
vdW (Å)	Damping (Å)	vdW	Damping	5CPV	1B8L	4ICB	4IHB	1ZOO	Acetate
2.90	1.39	2.90	1.65	1.3	2.3	2.0	4.0	0.2	0.9
3.59	1.39	3.59	1.65	2.6	4.0	3.4	3.7	1.9	2.0
2.90	1.65	3.59	1.65	-9.6	-5.4	-9.2	-7.3	0.1	-0.7
2.90	1.39	3.59	1.39	-10.8	-7.0	-10.6	-7.1	-1.6	-1.8

Table 19: Relative binding free energies (kcal/mol) between different combination of vdW and polarization parameters.

The damping length is defined by $\left(\frac{\alpha}{a}\right)^{\frac{1}{3}}$, where α is the polarizability and a is the damping parameter in AMOEBA.

4. BENCHMARK OF GPW METHOD

Index	Name	TPSS-D3/def2TZVPP/psi4	TPSS-D3/GPW/cp2k	Reference CCSD(T)
1	HB-1	-3.53	-3.14	-3.13
2	HB-2	-5.62	-5.15	-4.99
3	HB-3	-20.20	-19.86	-18.75
4	HB-4	-16.61	-16.61	-16.06
5	HB-5	-20.88	-20.43	-20.64
6	HB-6	-18.26	-18.02	-16.93
7	HB-7	-17.34	-16.86	-16.66
8	DD-1	-0.62	-0.60	-0.53
9	DD-2	-1.37	-1.44	-1.47
10	DD-3	-1.45	-1.52	-1.45
11	DD-4	-2.99	-3.18	-2.65
12	DD-6	-4.05	-4.15	-4.26
13	MX-5	-9.40	-9.51	-9.81
14	DD-7	-4.77	-5.00	-4.52
15	MX-8	-10.95	-11.06	-11.73
16	MX-1	-1.78	-1.84	-1.50
17	MX-2	-3.86	-3.39	-3.28
18	MX-3	-2.56	-2.46	-2.31
19	MX-4	-4.73	-4.82	-4.54
20	DD-5	-2.84	-2.90	-2.72
21	MX-6	-5.81	-5.82	-5.63
22	MX-7	-6.94	-6.70	-7.10
RMSE from psi4			0.25	
RMSE from Ref		0.56	0.44	

CP2K used GPW with MOLOPT-TZV2P + PW (400 Ry cutoff) and GTH-PP, MT Poisson solver.
RMSEs of other methods in psi4 for comparison (in kcal/mol). ωB97X-D/def2TZVPD, 0.21; ωB97X-V/def2TZVPD, 0.36; RI-MP2/def2QZVPPD, 1.21.

Table 20: Performance of psi4 and cp2k on S22 benchmark (energy in kcal/mol).

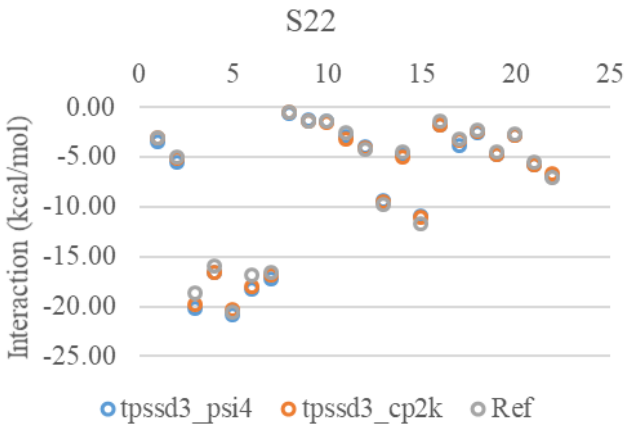


Figure 53: Comparison of psi4 and cp2k on S22 benchmark.

5. COMPARISON OF DFT'S AND AMOEBA+

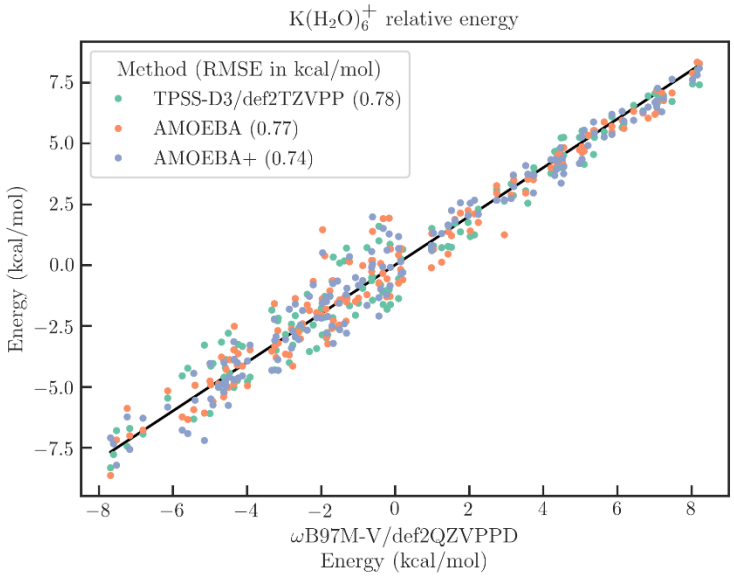


Figure 54: Accuracy of AMOEBA and DFT on $K(H_2O)_6^+$ conformational energies compared to ωB97M-V/def2QZVPPD.

The structures were taken from TPSS-D3/GPW simulation at 300 K.

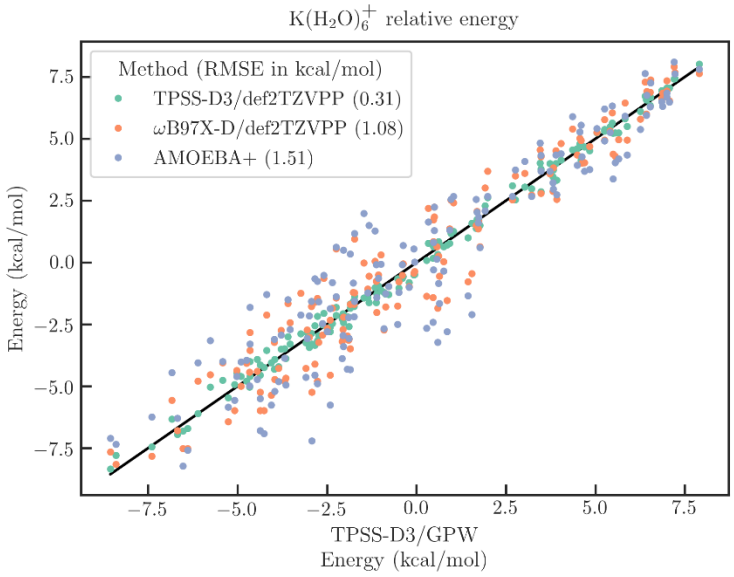


Figure 55: Deviation of AMOEBA and Gaussian DFT's from TPSS-D3/GPW on $K(H_2O)_6^+$ conformational energies.

The GPW calculation in CP2K used the wavelet Poisson solver. Small deviations imply small statistical errors in the free energy calculation.

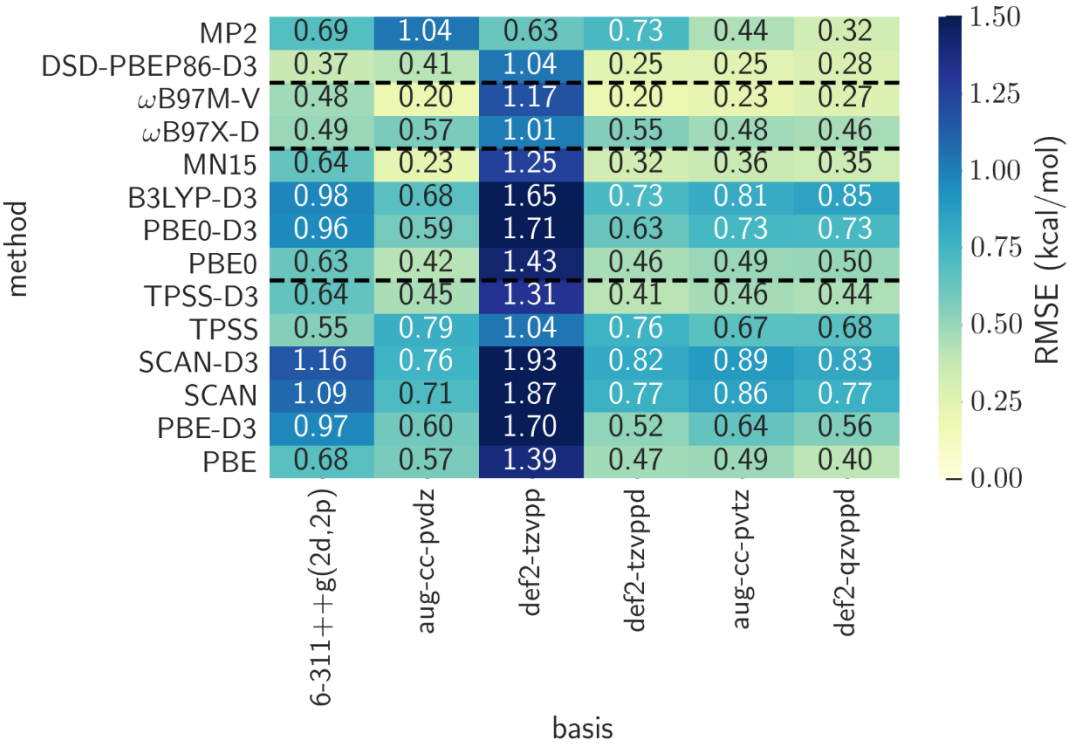


Figure 56: Comparison of QM methods on X-H_2O (X=Li, Na, K, Rb, Cs, F, Cl, Br, I) dimers.

For all the alkali metals and elements where the designated basis set is not available, the def2-tzvpp or def2-qzvpp basis set is used. The reference energy is calculated by MP2/CBS (aTZ and aQZ extrapolation) + δCCSD(T)/aTZ, which is consistent with the literature. The structures for X=Li, Na, K, F, Cl were taken from literature.

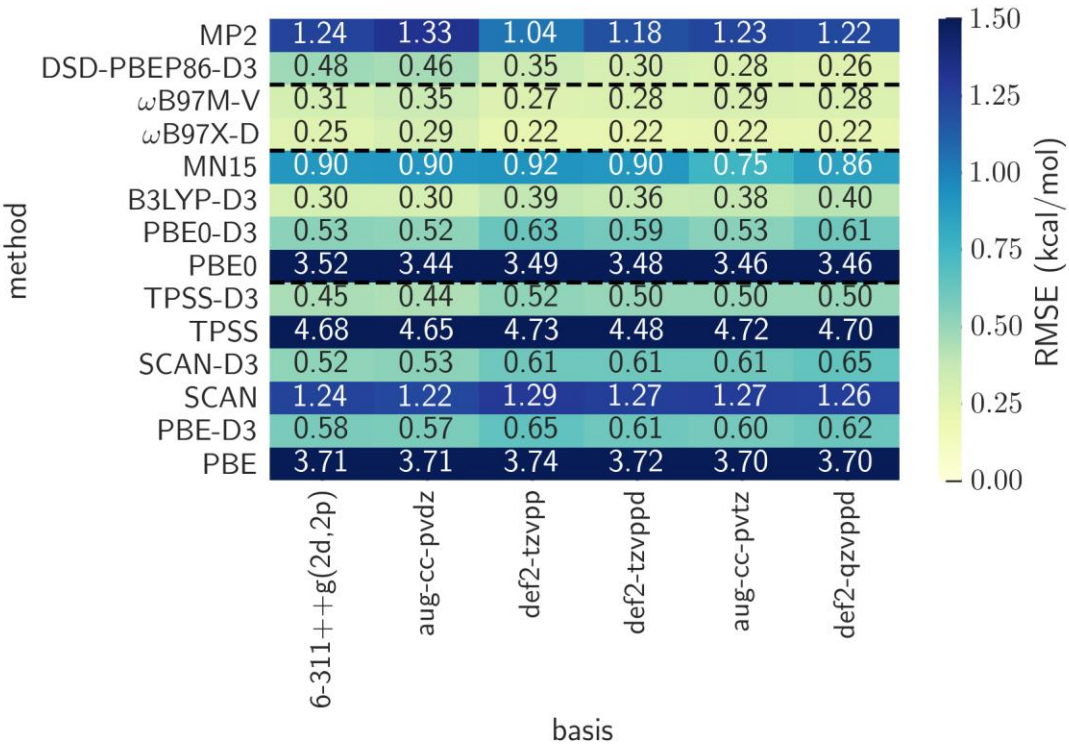


Figure 57: Comparison of QM methods on S22 data set (biologically relevant dimers).

The reference energy is taken from the literature, which was calculated by CCSD(T)/CBS or MP2/CBS + δCCSD(T).

6. STRUCTURES OF ION-WATER CLUSTERS IN QCT CALCULATION

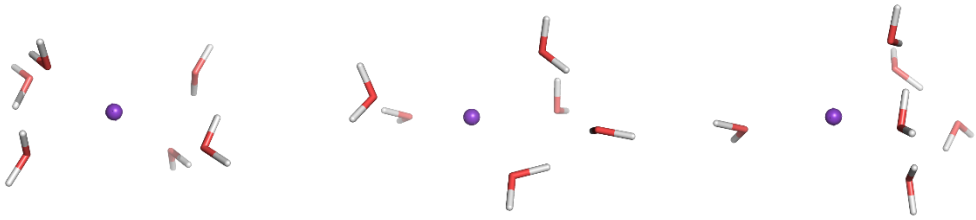


Figure 58: Effect of cluster structure on QM harmonic results.

ΔG_{hyd}(kcal/mol) calculated by TPSS-D3/def2TZVPP from left to right, -80.4, -81.9, -82.6.

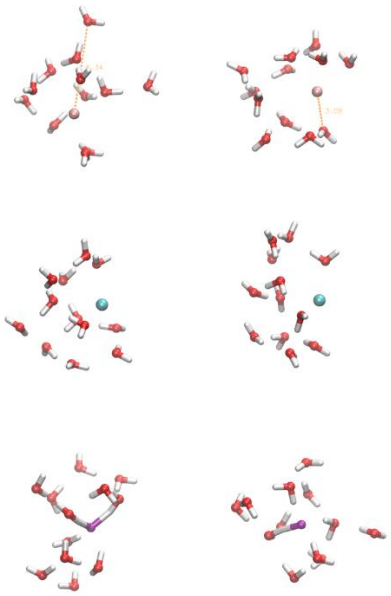


Figure 59: Structures of K, Rb, F clusters (from top to bottom) from AIMD simulations.

7. BENCHMARK OF QM METHODS FOR NONCOVALENT INTERACTIONS

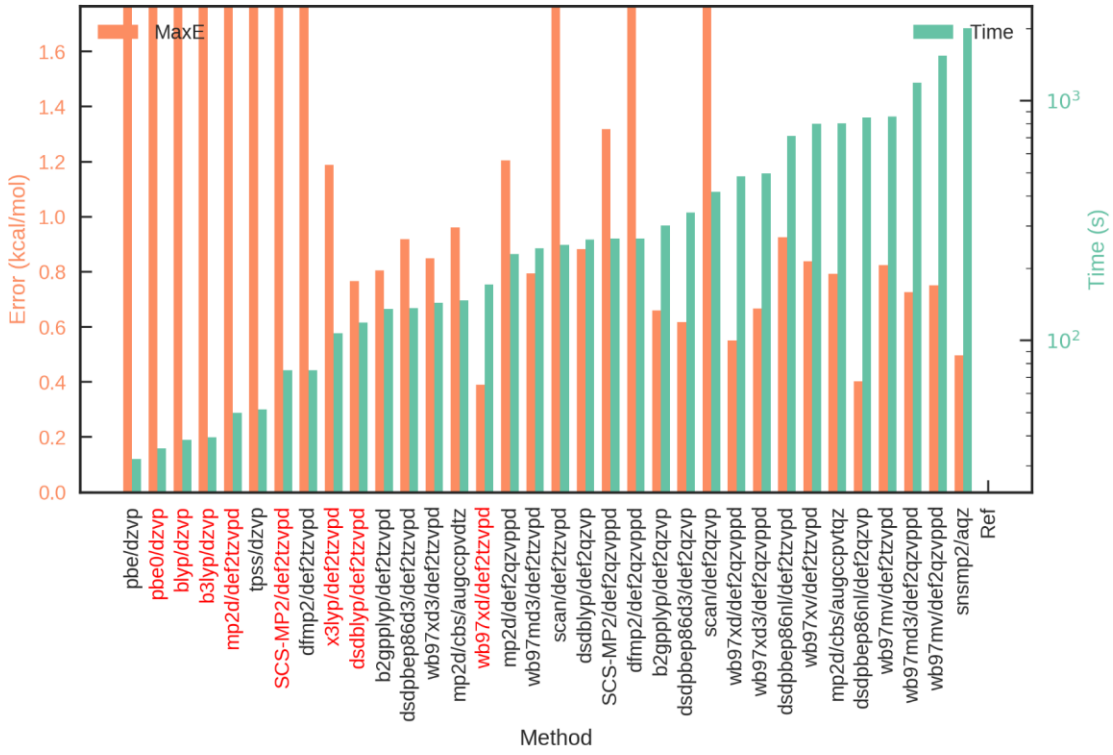


Figure 60: Benchmark of QM methods for the S22 dataset.

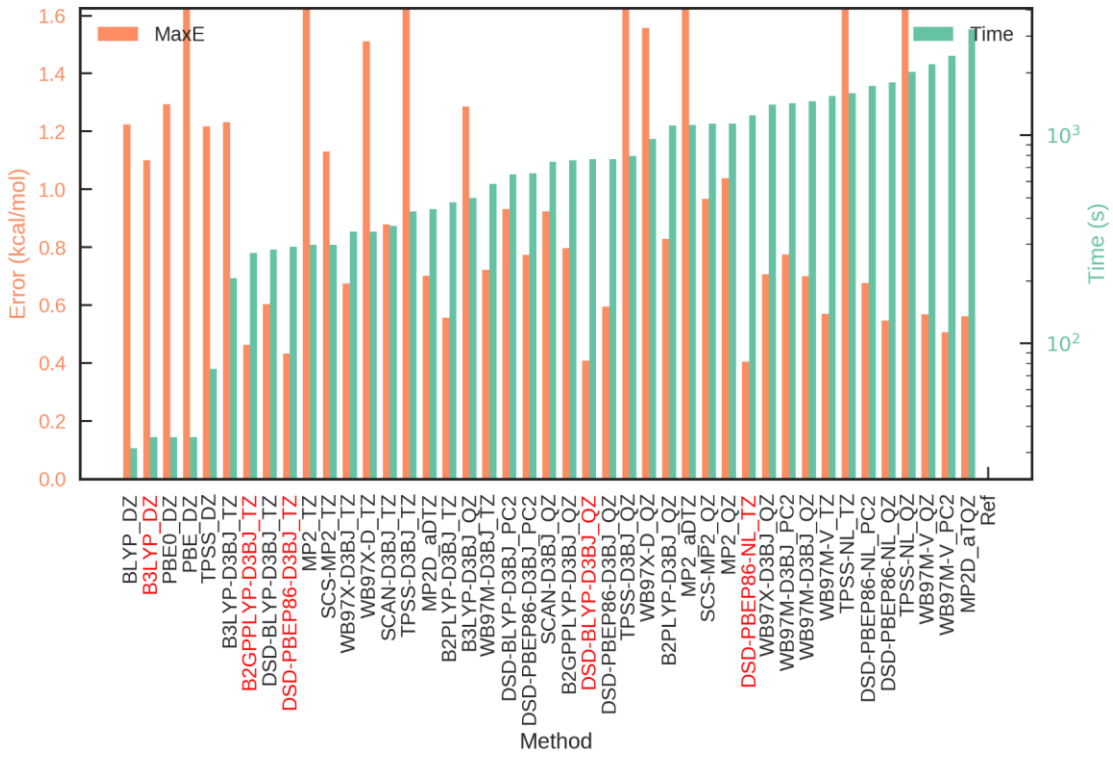


Figure 61: Benchmark of QM methods for the PCONF dataset.

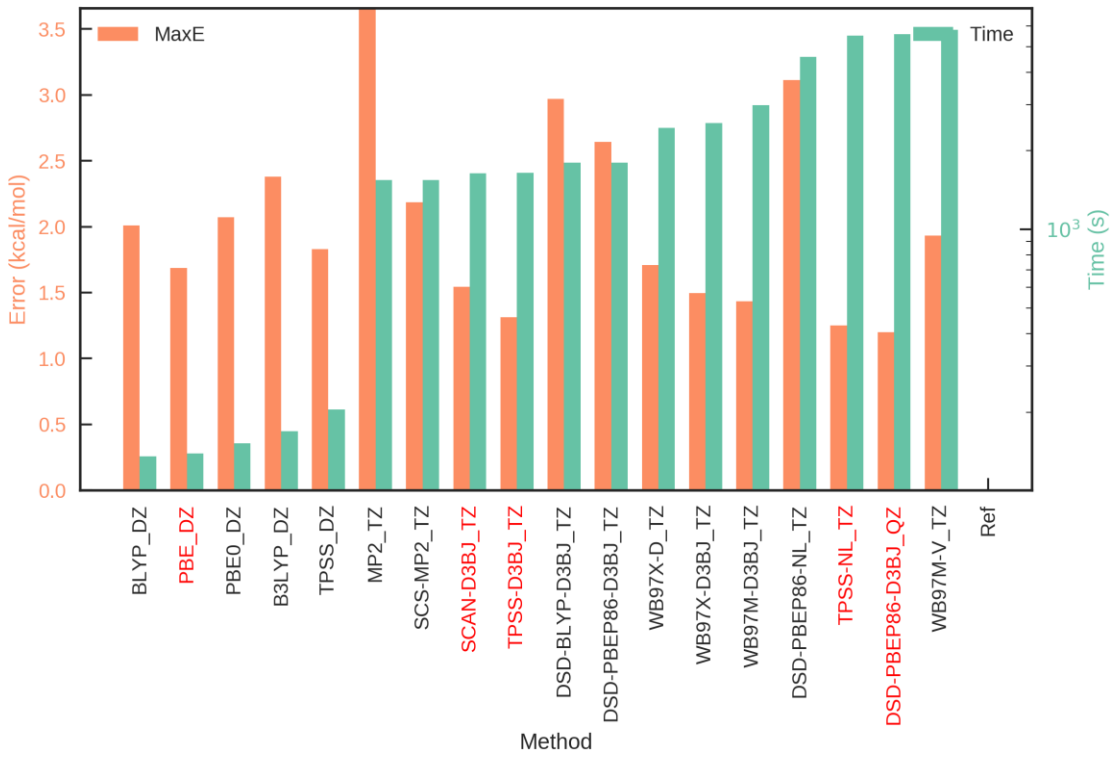


Figure 62: Benchmark of QM methods for the UPU46 dataset.

Bibliography

1 Jing, Z. *et al.* Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics* **48**, 371-394, doi:10.1146/annurev-biophys-070317-033349 (2019).

2 Ruiz Pestana, L., Mardirossian, N., Head-Gordon, M. & Head-Gordon, T. Ab initio molecular dynamics simulations of liquid water using high quality meta-GGA functionals. *Chemical Science* **8**, 3554-3565, doi:10.1039/C6SC04711D (2017).

3 Hagler, A. T. Force field development phase II: Relaxation of physics-based criteria… or inclusion of more rigorous physics into the representation of molecular energetics. *J. Comput. Aided Mol. Des.* **33**, 205-264, doi:10.1007/s10822-018-0134-x (2019).

4 Dauber-Osguthorpe, P. & Hagler, A. T. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *J. Comput. Aided Mol. Des.* **33**, 133-203, doi:10.1007/s10822-018-0111-4 (2019).

5 Sun, H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *The Journal of Physical Chemistry B* **102**, 7338-7364, doi:10.1021/jp980939v (1998).

6 Mortier, J. *et al.* The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes. *Drug Discovery Today* **20**, 686-702, doi:https://doi.org/10.1016/j.drudis.2015.01.003 (2015).

7 Goh, B. C. *et al.* Computational Methodologies for Real-Space Structural Refinement of Large Macromolecular Complexes. *Annual Review of Biophysics* **45**, 253-278, doi:10.1146/annurev-biophys-062215-011113 (2016).

8 Walsh, T. R. & Knecht, M. R. Biointerface Structural Effects on the Properties and Applications of Bioinspired Peptide-Based Nanomaterials. *Chem. Rev.* **117**, 12641-12704, doi:10.1021/acs.chemrev.7b00139 (2017).

9 Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878-3888, doi:10.1021/ct400314y (2013).

10 Carsten, K. *et al.* Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *J. Comput. Chem.* **36**, 1990-2008, doi:doi:10.1002/jcc.24030 (2015).

11 Eastman, P. *et al.* OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *bioRxiv*, doi:10.1101/091801 (2016).

12 Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386-2396, doi:10.1021/jacs.7b12191 (2018).

13 Faradjian, A. K. & Elber, R. Computing time scales from reaction coordinates by milestoning. *The Journal of Chemical Physics* **120**, 10880-10889, doi:10.1063/1.1738640 (2004).

14 Elber, R. Perspective: Computer simulations of long time dynamics. *The Journal of Chemical Physics* **144**, 060901, doi:10.1063/1.4940794 (2016).
15 Omar, D., Lee‐Ping, W. & Teresa, H. G. Advanced models for water simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8**, e1355, doi:doi:10.1002/wcms.1355 (2018).
16 Lemkul, J. A., Huang, J., Roux, B. & MacKerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chemical Reviews* **116**, 4983-5013, doi:10.1021/acs.chemrev.5b00505 (2016).
17 Ren, P. *et al.* Biomolecular electrostatics and solvation: a computational perspective. *Q. Rev. Biophys.* **45**, 427-491, doi:10.1017/s003358351200011x (2012).
18 Cisneros, G. A., Karttunen, M., Ren, P. & Sagui, C. Classical Electrostatics for Biomolecular Simulations. *Chem. Rev.* **114**, 779-814, doi:10.1021/cr300461d (2014).
19 Baker, C. M. Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **5**, 241-254, doi:doi:10.1002/wcms.1215 (2015).
20 Marshall, G. R. Limiting assumptions in molecular modeling: electrostatics. *J. Comput. Aided Mol. Des.* **27**, 107-114, doi:10.1007/s10822-013-9634-x (2013).
21 Jakobsen, S. & Jensen, F. Systematic Improvement of Potential-Derived Atomic Multipoles and Redundancy of the Electrostatic Parameter Space. *J. Chem. Theory Comput.* **10**, 5493-5504, doi:10.1021/ct500803r (2014).
22 Kolář, M. H. & Hobza, P. Computer Modeling of Halogen Bonds and Other σ-Hole Interactions. *Chem. Rev.* **116**, 5155-5187, doi:10.1021/acs.chemrev.5b00560 (2016).
23 Unke, O. T., Devereux, M. & Meuwly, M. Minimal distributed charges: Multipolar quality at the cost of point charge electrostatics. *The Journal of Chemical Physics* **147**, 161712, doi:10.1063/1.4993424 (2017).
24 Kramer, C., Spinn, A. & Liedl, K. R. Charge Anisotropy: Where Atomic Multipoles Matter Most. *J. Chem. Theory Comput.* **10**, 4488-4496, doi:10.1021/ct5005565 (2014).
25 Gresh, N., Cisneros, G. A., Darden, T. A. & Piquemal, J.-P. Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand−Macromolecule Complexes. A Bottom-Up Strategy. *J. Chem. Theory Comput.* **3**, 1960-1986, doi:10.1021/ct700134r (2007).
26 Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B* **114**, 2549-2564, doi:10.1021/jp910674d (2010).
27 Halgren, T. A. & Damm, W. Polarizable force fields. *Curr. Opin. Struct. Biol.* **11**, 236-242, doi:Doi 10.1016/S0959-440x(00)00196-2 (2001).
28 Warshel, A., Kato, M. & Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **3**, 2034-2045, doi:10.1021/ct700127w (2007).

29 Piotr, C., François-Yves, D., Yong, D. & Junmei, W. Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **21**, 333102 (2009).
30 Wang, J., Wolf Romain, M., Caldwell James, W., Kollman Peter, A. & Case David, A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157-1174, doi:10.1002/jcc.20035 (2004).
31 MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* **102**, 3586-3616, doi:10.1021/jp973084f (1998).
32 Debiec, K. T. *et al.* Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **12**, 3926-3947, doi:10.1021/acs.jctc.6b00567 (2016).
33 Cerutti, D. S., Rice, J. E., Swope, W. C. & Case, D. A. Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *The Journal of Physical Chemistry B* **117**, 2328-2338, doi:10.1021/jp311851r (2013).
34 Swope, W. C., Horn, H. W. & Rice, J. E. Accounting for Polarization Cost When Using Fixed Charge Force Fields. I. Method for Computing Energy. *The Journal of Physical Chemistry B* **114**, 8621-8630, doi:10.1021/jp911699p (2010).
35 Misquitta, A. J., Stone, A. J. & Fazeli, F. Distributed Multipoles from a Robust Basis-Space Implementation of the Iterated Stockholder Atoms Procedure. *J. Chem. Theory Comput.* **10**, 5405-5418, doi:10.1021/ct5008444 (2014).
36 Stone, A. *The Theory of Intermolecular Forces*. (Oxford University Press, USA, 2016).
37 Sandeep, P. & L., B. C. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **25**, 1-16, doi:doi:10.1002/jcc.10355 (2004).
38 Huang, J., Simmonett, A. C., Pickard, F. C., MacKerell, A. D. & Brooks, B. R. Mapping the Drude polarizable force field onto a multipole and induced dipole model. *The Journal of Chemical Physics* **147**, 161702, doi:10.1063/1.4984113 (2017).
39 Chen, J. & Martínez, T. J. QTPIE: Charge transfer with polarization current equalization. A fluctuating charge model with correct asymptotics. *Chem. Phys. Lett.* **438**, 315-320, doi:https://doi.org/10.1016/j.cplett.2007.02.065 (2007).
40 Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024-10035, doi:10.1021/ja00051a040 (1992).
41 Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *Npj Computational Materials* **2**, 15011, doi:10.1038/npjcompumats.2015.11 (2016).
42 Stern, H. A., Rittner, F., Berne, B. J. & Friesner, R. A. Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential

function. *The Journal of Chemical Physics* **115**, 2237-2251, doi:10.1063/1.1376165 (2001).
43 Mei, Y. *et al.* Numerical Study on the Partitioning of the Molecular Polarizability into Fluctuating Charge and Induced Atomic Dipole Contributions. *The Journal of Physical Chemistry A* **119**, 5865-5882, doi:10.1021/acs.jpca.5b03159 (2015).
44 Leontyev, I. V. & Stuchebrukhov, A. A. Polarizable molecular interactions in condensed phase and their equivalent nonpolarizable models. *The Journal of Chemical Physics* **141**, 014103, doi:10.1063/1.4884276 (2014).
45 Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J Chem Phys* **147**, 161721, doi:10.1063/1.4999905 (2017).
46 Mao, Y., Demerdash, O., Head-Gordon, M. & Head-Gordon, T. Assessing Ion-Water Interactions in the AMOEBA Force Field Using Energy Decomposition Analysis of Electronic Structure Calculations. *J. Chem. Theory Comput.* **12**, 5422-5437, doi:10.1021/acs.jctc.6b00764 (2016).
47 Liu, C., Qi, R., Wang, Q., Piquemal, J.-P. & Ren, P. Capturing Many-body Interactions with Classical Dipole Induction Models. *J. Chem. Theory Comput.* **13**, 2751–2761 (2017).
48 Grossfield, A., Ren, P. & Ponder, J. W. Ion Solvation Thermodynamics from Simulation with a Polarizable Force Field. *J. Am. Chem. Soc.* **125**, 15671-15682, doi:10.1021/ja037005r (2003).
49 Jiao, D., Golubkov, P. A., Darden, T. A. & Ren, P. Calculation of protein–ligand binding free energy by using a polarizable potential. *Proceedings of the National Academy of Sciences* **105**, 6290-6295, doi:10.1073/pnas.0711686105 (2008).
50 Verstraelen, T., Vandenbrande, S. & Ayers, P. W. Direct computation of parameters for accurate polarizable force fields. *The Journal of Chemical Physics* **141**, 194114, doi:10.1063/1.4901513 (2014).
51 Wang, L.-P. *et al.* Systematic Improvement of a Classical Molecular Model of Water. *The Journal of Physical Chemistry B* **117**, 9956-9972, doi:10.1021/jp403802c (2013).
52 Tazi, S. *et al.* A transferable ab initio based force field for aqueous ions. *The Journal of Chemical Physics* **136**, 114507, doi:10.1063/1.3692965 (2012).
53 Li, Y. *et al.* Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **13**, 4492-4503, doi:10.1021/acs.jctc.7b00521 (2017).
54 Wang, J. *et al.* Development of Polarizable Models for Molecular Mechanical Calculations. 4. van der Waals Parametrization. *The Journal of Physical Chemistry B* **116**, 7088-7101, doi:10.1021/jp3019759 (2012).
55 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *Journal of Physical Chemistry B* **107**, 5933-5947, doi:10.1021/jp027815+ (2003).

56 Zhang, C. *et al.* AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **14**, 2084-2108, doi:10.1021/acs.jctc.7b01169 (2018).
57 Boresch, S., Tettinger, F., Leitgeb, M. & Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B* **107**, 9535-9551, doi:10.1021/jp0217839 (2003).
58 Bell, D. R. *et al.* Calculating binding free energies of host–guest systems using the AMOEBA polarizable force field. *Phys. Chem. Chem. Phys.* **18**, 30261-30269, doi:10.1039/C6CP02509A (2016).
59 Liu, C., Qi, R., Wang, Q., Piquemal, J. P. & Ren, P. Capturing Many-Body Interactions with Classical Dipole Induction Models. *J. Chem. Theory Comput.*, doi:10.1021/acs.jctc.7b00225 (2017).
60 Rackers, J. A. *et al.* An optimized charge penetration model for use with the AMOEBA force field. *Phys. Chem. Chem. Phys.* **19**, 276-291 (2017).
61 Wang, Q. *et al.* General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *J. Chem. Theory Comput.* **11**, 2609-2618, doi:10.1021/acs.jctc.5b00267 (2015).
62 Liu, C., Piquemal, J.-P. & Ren, P. AMOEBA+ Classical Potential for Modeling Molecular Interactions. *J. Chem. Theory Comput.* **15**, 4122-4139, doi:10.1021/acs.jctc.9b00261 (2019).
63 Yang, X., Liu, C., Walker, B. D. & Ren, P. Accurate description of molecular dipole surface with charge flux implemented for molecular mechanics. *The Journal of Chemical Physics* **153**, 064103, doi:10.1063/5.0016376 (2020).
64 Liu, C., Piquemal, J.-P. & Ren, P. Implementation of Geometry-Dependent Charge Flux into the Polarizable AMOEBA+ Potential. *The Journal of Physical Chemistry Letters* **11**, 419-426, doi:10.1021/acs.jpclett.9b03489 (2020).
65 Mao, Y. *et al.* From Intermolecular Interaction Energies and Observable Shifts to Component Contributions and Back Again: A Tale of Variational Energy Decomposition Analysis. *Annu. Rev. Phys. Chem.*, doi:10.1146/annurev-physchem-090419-115149 (2021).
66 Misquitta, A. J. Charge Transfer from Regularized Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **9**, 5313-5326, doi:10.1021/ct400704a (2013).
67 Lao, K. U., Schäffer, R., Jansen, G. & Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **11**, 2473-2486, doi:10.1021/ct5010593 (2015).
68 Wang, Z. *Polarizable Force Field Development, and Applications to Conformational Sampling and Free Energy Calculation* Ph. D. thesis, Washington University in St. Louis, (2018).
69 Clementi, E. & Raimondi, D. L. Atomic Screening Constants from SCF Functions. *The Journal of Chemical Physics* **38**, 2686-2689, doi:10.1063/1.1733573 (1963).

70 Schmid, R., Miah, A. M. & Sapunov, V. N. A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii). *Phys Chem Chem Phys* **2**, 97-102, doi:10.1039/A907160A (2000).
71 Marcus, Y. Thermodynamics of solvation of ions. Part 5.-Gibbs free energy of hydration at 298.15 K. *J Chem Soc Faraday Trans* **87**, 2995-2999, doi:10.1039/FT9918702995 (1991).
72 Tissandier, M. D. *et al.* The Proton's Absolute Aqueous Enthalpy and Gibbs Free Energy of Solvation from Cluster-Ion Solvation Data. *The Journal of Physical Chemistry A* **102**, 7787-7794, doi:10.1021/jp982638r (1998).
73 Jing, Z., Qi, R., Liu, C. & Ren, P. Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field. *J. Chem. Phys.* **147**, 161733 (2017).
74 Dudev, T. & Lim, C. Competition among Metal Ions for Protein Binding Sites: Determinants of Metal Ion Selectivity in Proteins. *Chem. Rev.* **114**, 538-556, doi:10.1021/cr4004665 (2014).
75 Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. Metal ions in biological catalysis: from enzyme databases to general principles. *JBIC Journal of Biological Inorganic Chemistry* **13**, 1205-1218, doi:10.1007/s00775-008-0404-5 (2008).
76 Fukada, T., Yamasaki, S., Nishida, K., Murakami, M. & Hirano, T. Zinc homeostasis and signaling in health and diseases. *JBIC Journal of Biological Inorganic Chemistry* **16**, 1123-1134, doi:10.1007/s00775-011-0797-4 (2011).
77 Berridge, M. J., Lipp, P. & Bootman, M. D. The versatility and universality of calcium signalling. *Nature Reviews Molecular Cell Biology* **1**, 11, doi:10.1038/35036035 (2000).
78 Roderick, H. L. & Cook, S. J. Ca2+ signalling checkpoints in cancer: remodelling Ca2+ for cancer cell proliferation and survival. *Nature Reviews Cancer* **8**, 361, doi:10.1038/nrc2374 (2008).
79 de Baaij, J. H. F., Hoenderop, J. G. J. & Bindels, R. J. M. Magnesium in Man: Implications for Health and Disease. *Physiol. Rev.* **95**, 1-46, doi:10.1152/physrev.00012.2014 (2015).
80 Lagardere, L. *et al.* Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chemical Science* **9**, 956-972, doi:10.1039/C7SC04531J (2018).
81 Zhang, J., Yang, W., Piquemal, J. P. & Ren, P. Modeling Structural Coordination and Ligand Binding in Zinc Proteins with a Polarizable Potential. *Journal of Chemical Theory and Computation* **8**, 1314-1324 PMCID: PMC3383645 (2012).
82 Wu, J. C., Piquemal, J. P., Chaudret, R., Reinhardt, P. & Ren, P. Y. Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *Journal of Chemical Theory and Computation* **6**, 2059-2070 PMCID: PMC2992432 (2010).

83 Cornell, W. D. *et al.* A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).
84 Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688, doi:10.1002/jcc.20290 (2005).
85 Brooks, B. R. *et al.* Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comput Chem* **4**, 187-217, doi:Doi 10.1002/Jcc.540040211 (1983).
86 Jorgensen, W. L., Maxwell, D. S. & TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225-11236 (1996).
87 Reif, M. M., Hünenberger, P. H. & Oostenbrink, C. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *J. Chem. Theory Comput.* **8**, 3705-3723, doi:10.1021/ct300156h (2012).
88 Ren, P. *et al.* Biomolecular electrostatics and solvation: a computational perspective. *Q. Rev. Biophys.* **45**, 427-491, doi:10.1017/s003358351200011x (2012).
89 Shi, Y., Ren, P., Schnieders, M. & Piquemal, J.-P. in *Reviews in Computational Chemistry Volume 28* 51-86 (John Wiley & Sons, Inc, 2015).
90 Huang, J., Lopes, P. E. M., Roux, B. & MacKerell, A. D. Recent Advances in Polarizable Force Fields for Macromolecules: Microsecond Simulations of Proteins Using the Classical Drude Oscillator Model. *J. Phys. Chem. Lett.* **5**, 3144-3150, doi:10.1021/jz501315h (2014).
91 Cisneros, G. A., Karttunen, M., Ren, P. & Sagui, C. Classical Electrostatics for Biomolecular Simulations. *Chem. Rev.* **114**, 779-814, doi:10.1021/cr300461d (2014).
92 Engkvist, O., Astrand, P. O. & Karlstrom, G. Accurate intermolecular potentials obtained from molecular wave functions: Bridging the gap between quantum chemistry and molecular simulations. *Chem. Rev.* **100**, 4087-4108, doi:10.1021/cr9900477 (2000).
93 Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).
94 Liu, C., Qi, R., Wang, Q., Piquemal, J. P. & Ren, P. Capturing Many-Body Interactions with Classical Dipole Induction Models. *J Chem Theory Comput*, doi:10.1021/acs.jctc.7b00225 (2017).
95 Kumar, R., Wang, F.-F., Jenness, G. R. & Jordan, K. D. A second generation distributed point polarizable water model. *J. Chem. Phys.* **132**, 014309, doi:doi:http://dx.doi.org/10.1063/1.3276460 (2010).
96 Christie, R. A. & Jordan, K. D. in *Intermolecular Forces and Clusters II* (ed D. J. Wales) 27-41 (Springer Berlin Heidelberg, 2005).
97 Ahlstrand, E., Hermansson, K. & Friedman, R. Interaction Energies in Complexes of Zn and Amino Acids: A Comparison of Ab Initio and Force Field Based

Calculations. *J Phys Chem A* **121**, 2643-2654, doi:10.1021/acs.jpca.6b12969 (2017).
98 Gresh, N. *et al.* Polarizable Molecular Mechanics Studies of Cu(I)/Zn(II) Superoxide Dismutase: Bimetallic Binding Site and Structured Waters. *J Comput Chem* **35**, 2096-2106, doi:10.1002/jcc.23724 (2014).
99 Suarez, D., Rayon, V. M., Diaz, N. & Valdes, H. Ab Initio Benchmark Calculations on Ca(II) Complexes and Assessment of Density Functional Theory Methodologies. *J Phys Chem A* **115**, 11331-11343, doi:10.1021/jp205101z (2011).
100 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *J. Chem. Phys.* **140**, doi:10.1063/1.4867135 (2014).
101 McDaniel, J. G. & Schmidt, J. R. Next-Generation Force Fields from Symmetry-Adapted Perturbation Theory. *Annu. Rev. Phys. Chem.* **67**, 467-488, doi:10.1146/annurev-physchem-040215-112047 (2016).
102 Harger, M. *et al.* Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *J. Comput. Chem.* **38**, 2047-2055, doi:10.1002/jcc.24853 (2017).
103 Mardirossian, N. & Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110, doi:10.1063/1.4952647 (2016).
104 Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J. Chem. Phys.* **147**, 161721, doi:10.1063/1.4999905 (2017).
105 Lao, K. U. & Herbert, J. M. Energy Decomposition Analysis with a Stable Charge-Transfer Term for Interpreting Intermolecular Interactions. *J. Chem. Theory Comput.* **12**, 2569-2582, doi:10.1021/acs.jctc.6b00155 (2016).
106 Jing, Z., Liu, C., Qi, R. & Ren, P. Many-body effect determines the selectivity for Ca²⁺ and Mg²⁺ in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7495-E7501, doi:10.1073/pnas.1805049115 (2018).
107 Yang, W. *et al.* Rational Design of a Calcium-Binding Protein. *Journal of the American Chemical Society* **125**, 6165-6171, doi:10.1021/ja034724x (2003).
108 Laires, M. J., Monteiro, C. P. & Bicho, M. Role of cellular magnesium in health and human disease. *Front. Biosci.* **9**, 262-276 (2004).
109 Yu, H., Noskov, S. Y. & Roux, B. Two mechanisms of ion selectivity in protein binding sites. *Proceedings of the National Academy of Sciences* **107**, 20329-20334, doi:10.1073/pnas.1007150107 (2010).
110 Gouaux, E. & MacKinnon, R. Principles of Selective Ion Transport in Channels and Pumps. *Science* **310**, 1461-1465, doi:10.1126/science.1113666 (2005).
111 Dudev, T. & Lim, C. Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chem Rev* **114**, 538-556, doi:10.1021/cr4004665 (2014).

112 Gifford, Jessica L., Walsh, Michael P. & Vogel, Hans J. Structures and metal-ion-binding properties of the Ca²⁺-binding helix–loop–helix EF-hand motifs. *Biochem. J.* **405**, 199-221, doi:10.1042/bj20070255 (2007).
113 Cates, M. S. *et al.* Metal-ion affinity and specificity in EF-hand proteins: coordination geometry and domain plasticity in parvalbumin. *Structure* **7**, 1269-1278, doi:10.1016/S0969-2126(00)80060-X (1999).
114 Snyder, E. E., Buoscio, B. W. & Falke, J. J. Calcium(II) site specificity: effect of size and charge on metal ion binding to an EF-hand-like site. *Biochemistry* **29**, 3937-3943, doi:10.1021/bi00468a021 (1990).
115 Needham, J. V., Chen, T. Y. & Falke, J. J. Novel ion specificity of a carboxylate cluster magnesium(II) binding site: Strong charge selectivity and weak size selectivity. *Biochemistry* **32**, 3363-3367, doi:10.1021/bi00064a020 (1993).
116 Abdullah, N., Padmanarayana, M., Marty, Naomi J. & Johnson, Colin P. Quantitation of the Calcium and Membrane Binding Properties of the C2 Domains of Dysferlin. *Biophys. J.* **106**, 382-389, doi:10.1016/j.bpj.2013.11.4492 (2013).
117 Andersson, M. *et al.* Structural basis for the negative allostery between Ca2+- and Mg2+ -binding in the intracellular Ca2+ -receptor calbindin D9k. *Protein Sci.* **6**, 1139-1147, doi:10.1002/pro.5560060602 (1997).
118 Breukels, V., Konijnenberg, A., Nabuurs, S. M., Touw, W. G. & Vuister, G. W. The Second Ca2+-Binding Domain of NCX1 Binds Mg2+ with High Affinity. *Biochemistry* **50**, 8804-8812, doi:10.1021/bi201134u (2011).
119 Kherb, J., Flores, S. C. & Cremer, P. S. Role of Carboxylate Side Chains in the Cation Hofmeister Series. *The Journal of Physical Chemistry B* **116**, 7389-7397, doi:10.1021/jp212243c (2012).
120 Babu, C. S., Dudev, T., Casareno, R., Cowan, J. A. & Lim, C. A combined experimental and theoretical study of divalent metal ion selectivity and function in proteins: application to E. coli ribonuclease H1. *J Am Chem Soc* **125**, 9318-9328, doi:10.1021/ja034956w (2003).
121 Dudev, T., Lin, Dudev, M. & Lim, C. First−Second Shell Interactions in Metal Binding Sites in Proteins: A PDB Survey and DFT/CDM Calculations. *Journal of the American Chemical Society* **125**, 3168-3180, doi:10.1021/ja0209722 (2003).
122 Dudev, T. & Lim, C. Metal Binding Affinity and Selectivity in Metalloproteins: Insights from Computational Studies. *Annual Review of Biophysics* **37**, 97-116, doi:10.1146/annurev.biophys.37.032807.125811 (2008).
123 Babu, C. S., Dudev, T. & Lim, C. Differential Role of the Protein Matrix on the Binding of a Catalytic Aspartate to Mg2+ vs Ca2+: Application to Ribonuclease H. *Journal of the American Chemical Society* **135**, 6541-6548, doi:10.1021/ja4006095 (2013).
124 Kucharski, A. N., Scott, C. E., Davis, J. P. & Kekenes-Huskey, P. M. Understanding Ion Binding Affinity and Selectivity in β-Parvalbumin Using Molecular Dynamics and Mean Spherical Approximation Theory. *The Journal of Physical Chemistry B* **120**, 8617-8630, doi:10.1021/acs.jpcb.6b02666 (2016).

125 Cates, M. S., Teodoro, M. L. & Phillips, G. N. Molecular Mechanisms of Calcium and Magnesium Binding to Parvalbumin. *Biophysical Journal* **82**, 1133-1146, doi:https://doi.org/10.1016/S0006-3495(02)75472-6 (2002).

126 Lepšík, M. & Field, M. J. Binding of Calcium and Other Metal Ions to the EF-Hand Loops of Calmodulin Studied by Quantum Chemical Calculations and Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **111**, 10012-10022, doi:10.1021/jp0716583 (2007).

127 Ye, Y., Lee, H.-W., Yang, W., Shealy, S. & Yang, J. J. Probing Site-Specific Calmodulin Calcium and Lanthanide Affinity by Grafting. *Journal of the American Chemical Society* **127**, 3743-3750, doi:10.1021/ja042786x (2005).

128 Drake, S. K., Lee, K. L. & Falke, J. J. Tuning the Equilibrium Ion Affinity and Selectivity of the EF-Hand Calcium Binding Motif: Substitutions at the Gateway Position. *Biochemistry* **35**, 6697-6705, doi:10.1021/bi952430l (1996).

129 Dudev, T. & Lim, C. A DFT/CDM Study of Metal−Carboxylate Interactions in Metalloproteins: Factors Governing the Maximum Number of Metal-Bound Carboxylates. *Journal of the American Chemical Society* **128**, 1553-1561, doi:10.1021/ja055797e (2006).

130 Kohagen, M., Lepšík, M. & Jungwirth, P. Calcium Binding to Calmodulin by Molecular Dynamics with Effective Polarization. *The Journal of Physical Chemistry Letters* **5**, 3964-3969, doi:10.1021/jz502099g (2014).

131 Illingworth, C. J. R., Furini, S. & Domene, C. Computational Studies on Polarization Effects and Selectivity in K+ Channels. *Journal of Chemical Theory and Computation* **6**, 3780-3792, doi:10.1021/ct100276c (2010).

132 Sakharov, D. V. & Lim, C. Zn Protein Simulations Including Charge Transfer and Local Polarization Effects. *Journal of the American Chemical Society* **127**, 4921-4929, doi:10.1021/ja0429115 (2005).

133 Lu, Y., Mei, Y., Zhang, J. Z. H. & Zhang, D. Communications: Electron polarization critically stabilizes the Mg2+ complex in the catalytic core domain of HIV-1 integrase. *The Journal of Chemical Physics* **132**, 131101, doi:10.1063/1.3360769 (2010).

134 Zhekova, H. R., Ngo, V., da Silva, M. C., Salahub, D. & Noskov, S. Selective ion binding and transport by membrane proteins – A computational perspective. *Coordination Chemistry Reviews* **345**, 108-136, doi:https://doi.org/10.1016/j.ccr.2017.03.019 (2017).

135 Illingworth, C. J. & Domene, C. Many-body effects and simulations of potassium channels. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* **465**, 1701-1716, doi:10.1098/rspa.2009.0014 (2009).

136 Rao, L., Cui, Q. & Xu, X. Electronic Properties and Desolvation Penalties of Metal Ions Plus Protein Electrostatics Dictate the Metal Binding Affinity and Selectivity in the Copper Efflux Regulator. *Journal of the American Chemical Society* **132**, 18092-18102, doi:10.1021/ja103742k (2010).

137 Ngo, V. *et al.* Quantum Effects in Cation Interactions with First and Second Coordination Shell Ligands in Metalloproteins. *Journal of Chemical Theory and Computation* **11**, 4992-5001, doi:10.1021/acs.jctc.5b00524 (2015).
138 Jing, Z., Qi, R., Liu, C. & Ren, P. Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field. *J Chem Phys* **147**, 161733, doi:10.1063/1.4985921 (2017).
139 Kinraide, T. B. Improved scales for metal ion softness and toxicity. *Environmental Toxicology and Chemistry* **28**, 525-533, doi:10.1897/08-208.1 (2009).
140 Xu, H., Xu, D. C. & Wang, Y. Natural Indices for the Chemical Hardness/Softness of Metal Cations and Ligands. *ACS Omega* **2**, 7185-7193, doi:10.1021/acsomega.7b01039 (2017).
141 Henzl, M. T., Larson, J. D. & Agah, S. Estimation of parvalbumin Ca2+- and Mg2+-binding constants by global least-squares analysis of isothermal titration calorimetry data. *Analytical Biochemistry* **319**, 216-233, doi:https://doi.org/10.1016/S0003-2697(03)00288-4 (2003).
142 Vorup-Jensen, T., Waldron, T. T., Astrof, N., Shimaoka, M. & Springer, T. A. The connection between metal ion affinity and ligand affinity in integrin I domains. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1774**, 1148-1155, doi:https://doi.org/10.1016/j.bbapap.2007.06.014 (2007).
143 Christensen, T., Gooden, D. M., Kung, J. E. & Toone, E. J. Additivity and the Physical Basis of Multivalency Effects: A Thermodynamic Investigation of the Calcium EDTA Interaction. *Journal of the American Chemical Society* **125**, 7357-7366, doi:10.1021/ja021240c (2003).
144 Richard, R. M., Lao, K. U. & Herbert, J. M. Aiming for Benchmark Accuracy with the Many-Body Expansion. *Accounts of Chemical Research* **47**, 2828-2836, doi:10.1021/ar500119q (2014).
145 Harger, M. *et al.* Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *J Comput Chem* **38**, 2047-2055, doi:10.1002/jcc.24853 (2017).
146 Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *J Phys Chem B* **114**, 2549-2564, doi:10.1021/jp910674d (2010).
147 Shi, Y. *et al.* The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *J Chem Theory Comput* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).
148 Shi, Y. & Beck, T. L. Absolute ion hydration free energy scale and the surface potential of water via quantum simulation. *Proceedings of the National Academy of Sciences* **117**, 30151-30158, doi:10.1073/pnas.2017214117 (2020).
149 Sabo, D., Jiao, D., Varma, S., Pratt, L. R. & Rempe, S. B. Case study of Rb+(aq), quasi-chemical theory of ion hydration, and the no split occupancies rule. *Annual Reports Section "C" (Physical Chemistry)* **109**, 266-278, doi:10.1039/C3PC90009F (2013).

150 Whitfield, T. W. *et al.* Theoretical Study of Aqueous Solvation of K+ Comparing ab Initio, Polarizable, and Fixed-Charge Models. *J. Chem. Theory Comput.* **3**, 2068-2082, doi:10.1021/ct700172b (2007).
151 Muralidharan, A., Pratt, L. R., Chaudhari, M. I. & Rempe, S. B. Quasi-chemical theory for anion hydration and specific ion effects: Cl-(aq) vs. F-(aq). *Chemical Physics Letters: X* **4**, 100037, doi:https://doi.org/10.1016/j.cpletx.2019.100037 (2019).
152 Goerigk, L. *et al.* A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **19**, 32184-32215, doi:10.1039/C7CP04913G (2017).
153 West, D. & Estreicher, S. K. First-Principles Calculations of Vibrational Lifetimes and Decay Channels: Hydrogen-Related Modes in Si. *Phys. Rev. Lett.* **96**, 115504, doi:10.1103/PhysRevLett.96.115504 (2006).
154 LeMasurier, M., Heginbotham, L. & Miller, C. Kcsa: It's a Potassium Channel. *J. Gen. Physiol.* **118**, 303-314, doi:10.1085/jgp.118.3.303 (2001).
155 Flood, E., Boiteux, C., Lev, B., Vorobyov, I. & Allen, T. W. Atomistic Simulations of Membrane Ion Channel Conduction, Gating, and Modulation. *Chem. Rev.* **119**, 7737-7832, doi:10.1021/acs.chemrev.8b00630 (2019).
156 Yu, H., Noskov, S. Y. & Roux, B. Two mechanisms of ion selectivity in protein binding sites. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20329-20334, doi:10.1073/pnas.1007150107 (2010).
157 Varma, S., Rogers, D. M., Pratt, L. R. & Rempe, S. B. Design principles for K+ selectivity in membrane transport. *J. Gen. Physiol.* **137**, 479-488, doi:10.1085/jgp.201010579 (2011).
158 Roux, B. *et al.* Ion selectivity in channels and transporters. *J. Gen. Physiol.* **137**, 415-426, doi:10.1085/jgp.201010577 (2011).
159 Chaudhari, M. I., Vanegas, J. M., Pratt, L. R., Muralidharan, A. & Rempe, S. B. Hydration Mimicry by Membrane Ion Channels. *Annu. Rev. Phys. Chem.* **71**, 461-484, doi:10.1146/annurev-physchem-012320-015457 (2020).
160 Lockless, S. W. Determinants of cation transport selectivity: Equilibrium binding and transport kinetics. *J. Gen. Physiol.* **146**, 3-13, doi:10.1085/jgp.201511371 (2015).
161 Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K+ channel–Fab complex at 2.0 Å resolution. *Nature* **414**, 43-48, doi:10.1038/35102009 (2001).
162 Morais-Cabral, J. H., Zhou, Y. & MacKinnon, R. Energetic optimization of ion conduction rate by the K+ selectivity filter. *Nature* **414**, 37-42, doi:10.1038/35102000 (2001).
163 Jensen, M. Ø. *et al.* Principles of conduction and hydrophobic gating in K⁺ channels. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5833-5838, doi:10.1073/pnas.0911691107 (2010).

164 Bernèche, S. & Roux, B. Energetics of ion conduction through the K+ channel. *Nature* **414**, 73-77, doi:10.1038/35102067 (2001).
165 Medovoy, D., Perozo, E. & Roux, B. Multi-ion free energy landscapes underscore the microscopic mechanism of ion selectivity in the KcsA channel. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1858**, 1722-1732, doi:https://doi.org/10.1016/j.bbamem.2016.02.019 (2016).
166 Köpfer, D. A. *et al.* Ion permeation in K⁺ channels occurs by direct Coulomb knock-on. *Science* **346**, 352-355, doi:10.1126/science.1254840 (2014).
167 Kopec, W. *et al.* Direct knock-on of desolvated ions governs strict ion selectivity in K+ channels. *Nat. Chem.* **10**, 813-820, doi:10.1038/s41557-018-0105-9 (2018).
168 Furini, S. & Domene, C. Atypical mechanism of conduction in potassium channels. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16074-16077, doi:10.1073/pnas.0903226106 (2009).
169 Oakes, V., Furini, S. & Domene, C. Insights into the Mechanisms of K+ Permeation in K+ Channels from Computer Simulations. *J. Chem. Theory Comput.* **16**, 794-799, doi:10.1021/acs.jctc.9b00971 (2020).
170 Langan, P. S. *et al.* Anomalous X-ray diffraction studies of ion transport in K+ channels. *Nature Communications* **9**, 4540, doi:10.1038/s41467-018-06957-w (2018).
171 Öster, C. *et al.* The conduction pathway of potassium channels is water free under physiological conditions. *Science Advances* **5**, eaaw6756, doi:10.1126/sciadv.aaw6756 (2019).
172 Tilegenova, C. *et al.* Structure, function, and ion-binding properties of a K⁺ channel stabilized in the 2,4-ion–bound configuration. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16829-16834, doi:10.1073/pnas.1901888116 (2019).
173 Duboué-Dijon, E., Javanainen, M., Delcroix, P., Jungwirth, P. & Martinez-Seara, H. A practical guide to biologically relevant molecular simulations with charge scaling for electronic polarization. *J. Chem. Phys.* **153**, 050901, doi:10.1063/5.0017775 (2020).
174 Sun, R.-N. & Gong, H. Simulating the Activation of Voltage Sensing Domain for a Voltage-Gated Sodium Channel Using Polarizable Force Field. *J. Phys. Chem. Lett.* **8**, 901-908, doi:10.1021/acs.jpclett.7b00023 (2017).
175 Klesse, G., Rao, S., Tucker, S. J. & Sansom, M. S. P. Induced Polarization in Molecular Dynamics Simulations of the 5-HT3 Receptor Channel. *J. Am. Chem. Soc.* **142**, 9415-9427, doi:10.1021/jacs.0c02394 (2020).
176 Rodgers, M. T. & Armentrout, P. B. Cationic Noncovalent Interactions: Energetics and Periodic Trends. *Chem. Rev.* **116**, 5642-5687, doi:10.1021/acs.chemrev.5b00688 (2016).
177 Zhou, M. & MacKinnon, R. A Mutant KcsA K+ Channel with Altered Conduction Properties and Selectivity Filter Ion Distribution. *J. Mol. Biol.* **338**, 839-846, doi:https://doi.org/10.1016/j.jmb.2004.03.020 (2004).

178 Liu, S. *et al.* Ion-binding properties of a K⁺ channel selectivity filter in different conformations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15096-15100, doi:10.1073/pnas.1510526112 (2015).
179 Montoya, E. *et al.* Differential binding of monovalent cations to KcsA: Deciphering the mechanisms of potassium channel selectivity. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1859**, 779-788, doi:https://doi.org/10.1016/j.bbamem.2017.01.014 (2017).
180 Lockless, S. W., Zhou, M. & MacKinnon, R. Structural and Thermodynamic Properties of Selective Ion Binding in a K+ Channel. *PLoS Biol.* **5**, e121, doi:10.1371/journal.pbio.0050121 (2007).
181 Priest, C. *et al.* in *Channelrhodopsin: Methods and Protocols* (ed Robert E. Dempski) 17-28 (Springer US, 2021).
182 Devaraneni, P. K. *et al.* Semisynthetic K⁺ channels show that the constricted conformation of the selectivity filter is not the C-type inactivated state. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15698-15703, doi:10.1073/pnas.1308699110 (2013).
183 Cordero-Morales, J. F. *et al.* Molecular determinants of gating at the potassium-channel selectivity filter. *Nat. Struct. Mol. Biol.* **13**, 311-318, doi:10.1038/nsmb1069 (2006).
184 Sauer, D. B., Zeng, W., Canty, J., Lam, Y. & Jiang, Y. Sodium and potassium competition in potassium-selective and non-selective channels. *Nature Communications* **4**, 2721, doi:10.1038/ncomms3721 (2013).
185 Derebe, M. G. *et al.* Tuning the ion selectivity of tetrameric cation channels by changing the number of ion binding sites. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 598-602, doi:10.1073/pnas.1013636108 (2011).
186 Labro, A. J., Cortes, D. M., Tilegenova, C. & Cuello, L. G. Inverted allosteric coupling between activation and inactivation gates in K⁺ channels. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5426-5431, doi:10.1073/pnas.1800559115 (2018).
187 Alcayaga, C., Cecchi, X., Alvarez, O. & Latorre, R. Streaming potential measurements in Ca2+-activated K+ channels from skeletal and smooth muscle. Coupling of ion and water fluxes. *Biophys. J.* **55**, 367-371, doi:10.1016/S0006-3495(89)82814-0 (1989).
188 Kratochvil, H. T. *et al.* Instantaneous ion configurations in the K⁺ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **353**, 1040-1044, doi:10.1126/science.aag1447 (2016).
189 Cuello, L. G. *et al.* Structural basis for the coupling between activation and inactivation gates in K+ channels. *Nature* **466**, 272-275, doi:10.1038/nature09136 (2010).
190 Li, J. *et al.* Chemical substitutions in the selectivity filter of potassium channels do not rule out constricted-like conformations for C-type inactivation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11145-11150, doi:10.1073/pnas.1706983114 (2017).

191 Li, J., Ostmeyer, J., Cuello, L. G., Perozo, E. & Roux, B. Rapid constriction of the selectivity filter underlies C-type inactivation in the KcsA potassium channel. *J. Gen. Physiol.* **150**, 1408-1420, doi:10.1085/jgp.201812082 (2018).
192 Boiteux, C., Posson, D. J., Allen, T. W. & Nimigean, C. M. Selectivity filter ion binding affinity determines inactivation in a potassium channel. *Proc. Natl. Acad. Sci. U. S. A.*, 202009624, doi:10.1073/pnas.2009624117 (2020).
193 Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).
194 Liu, C., Qi, R., Wang, Q., Piquemal, J. P. & Ren, P. Capturing Many-Body Interactions with Classical Dipole Induction Models. *J. Chem. Theory Comput.* **13**, 2751-2761, doi:10.1021/acs.jctc.7b00225 (2017).
195 Chu, H. *et al.* Polarizable atomic multipole-based force field for DOPC and POPE membrane lipids. *Mol. Phys.* **116**, 1037-1050, doi:10.1080/00268976.2018.1436201 (2018).
196 Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405-413, doi:10.1021/acs.jctc.5b00935 (2016).
197 Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25, doi:https://doi.org/10.1016/j.softx.2015.06.001 (2015).
198 Rackers, J. A. *et al.* Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **14**, 5273-5289, doi:10.1021/acs.jctc.8b00529 (2018).
199 Grossfield, A. *WHAM: the weighted histogram analysis method. Version 2.0.10.1*, <http://membrane.urmc.rochester.edu/wordpress/?page_id=126> (
200 Jing, Z., Qi, R., Thibonnier, M. & Ren, P. Molecular Dynamics Study of the Hybridization between RNA and Modified Oligonucleotides. *J. Chem. Theory Comput.* **15**, 6422-6432, doi:10.1021/acs.jctc.9b00519 (2019).
201 Smith, C. I. E. & Zain, R. Therapeutic Oligonucleotides: State of the Art. *Annu. Rev. Pharmacol. Toxicol.* **59**, 605-630, doi:10.1146/annurev-pharmtox-010818-021050 (2019).
202 Khvorova, A. & Watts, J. K. The chemical evolution of oligonucleotide therapies of clinical utility. *Nat. Biotechnol.* **35**, 238-248, doi:10.1038/nbt.3765 (2017).
203 Shen, X. & Corey, D. R. Chemistry, mechanism and clinical status of antisense oligonucleotides and duplex RNAs. *Nucleic Acids Res.* **46**, 1584-1600, doi:10.1093/nar/gkx1239 (2017).
204 Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116**, 281-297, doi:https://doi.org/10.1016/S0092-8674(04)00045-5 (2004).
205 Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature Reviews Drug Discovery* **16**, 203, doi:10.1038/nrd.2016.246
https://www.nature.com/articles/nrd.2016.246#supplementary-information (2017).

206 van Rooij, E. & Kauppinen, S. Development of microRNA therapeutics is coming of age. *EMBO Mol. Med.* **6**, 851-864, doi:10.15252/emmm.201100899 (2014).
207 Hammond, S. M. An overview of microRNAs. *Advanced Drug Delivery Reviews* **87**, 3-14, doi:https://doi.org/10.1016/j.addr.2015.05.001 (2015).
208 Ebert, M. S. & Sharp, P. A. MicroRNA sponges: Progress and possibilities. *RNA* **16**, 2043-2050, doi:10.1261/rna.2414110 (2010).
209 Yin, H. *et al.* MicroRNA-133 Controls Brown Adipose Determination in Skeletal Muscle Satellite Cells by Targeting Prdm16. *Cell Metab.* **17**, 210-224, doi:10.1016/j.cmet.2013.01.004 (2013).
210 Fu, T. *et al.* MicroRNA 34a Inhibits Beige and Brown Fat Formation in Obesity in Part by Suppressing Adipocyte Fibroblast Growth Factor 21 Signaling and SIRT1 Function. *Mol. Cell. Biol.* **34**, 4130-4142, doi:10.1128/MCB.00596-14 (2014).
211 Chen, Y. *et al.* Exosomal microRNA miR-92a concentration in serum reflects human brown fat activity. *Nature Communications* **7**, 11420, doi:10.1038/ncomms11420 (2016).
212 Shamsi, F., Zhang, H. & Tseng, Y.-H. MicroRNA Regulation of Brown Adipogenesis and Thermogenic Energy Expenditure. *Front. Endocrinol. (Lausanne)* **8**, 205, doi:10.3389/fendo.2017.00205 (2017).
213 Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* **46**, D303-D307, doi:10.1093/nar/gkx1030 (2018).
214 Boczkowska, M., Guga, P. & Stec, W. J. Stereodefined phosphorothioate analogues of DNA: relative thermodynamic stability of the model PS-DNA/DNA and PS-DNA/RNA complexes. *Biochemistry* **41**, 12483-12487 (2002).
215 Lan, W. *et al.* Structural investigation into physiological DNA phosphorothioate modification. *Sci. Rep.* **6**, 25737, doi:10.1038/srep25737 (2016).
216 Iwamoto, N. *et al.* Control of phosphorothioate stereochemistry substantially increases the efficacy of antisense oligonucleotides. *Nat. Biotechnol.* **35**, 845-851, doi:10.1038/nbt.3948 (2017).
217 Quijano, E. C. C., Bahal, R., Ricciardi, A. S., Saltzman, W. M. & Glazer, P. M. in *The Yale journal of biology and medicine.*
218 Wan, W. B. & Seth, P. P. The Medicinal Chemistry of Therapeutic Oligonucleotides. *J. Med. Chem.* **59**, 9645-9667, doi:10.1021/acs.jmedchem.6b00551 (2016).
219 Juliano, R. L. The delivery of therapeutic oligonucleotides. *Nucleic Acids Res.* **44**, 6518-6548, doi:10.1093/nar/gkw236 (2016).
220 Sen, S. & Nilsson, L. MD Simulations of Homomorphous PNA, DNA, and RNA Single Strands: Characterization and Comparison of Conformations and Dynamics. *J. Am. Chem. Soc.* **123**, 7414-7422, doi:10.1021/ja0032632 (2001).
221 Soliva, R., Sherer, E., Luque, F. J., Laughton, C. A. & Orozco, M. Molecular Dynamics Simulations of PNA·DNA and PNA·RNA Duplexes in Aqueous Solution. *J. Am. Chem. Soc.* **122**, 5997-6008, doi:10.1021/ja000259h (2000).

222 Hartmann, B., Bertrand, H. O. & Fermandjian, S. Sequence effects on energetic and structural properties of phosphorothioate DNA: a molecular modelling study. *Nucleic Acids Res.* **27**, 3342-3347, doi:10.1093/nar/27.16.3342 (1999).

223 Lind, K. E., Sherlin, L. D., Mohan, V., Griffey, R. H. & Ferguson, D. M. in *Molecular Modeling of Nucleic Acids* Vol. 682 (eds Neocles B. Leontis & John SantaLucia) 41-54 (American Chemical Society, 1997).

224 Šponer, J. *et al.* RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chem. Rev.* **118**, 4177-4338, doi:10.1021/acs.chemrev.7b00427 (2018).

225 Obliosca, J. M. *et al.* LNA Thymidine Monomer Enables Differentiation of the Four Single-Nucleotide Variants by Melting Temperature. *J. Am. Chem. Soc.* **139**, 7110-7116, doi:10.1021/jacs.7b03395 (2017).

226 Mukherjee, S. & Bhattacharyya, D. Effect of phosphorothioate chirality on the grooves of DNA double helices: A molecular dynamics study. *Biopolymers* **73**, 269-282, doi:10.1002/bip.10550 (2004).

227 Zhang, Y.-C. *et al.* Theoretical Study on Steric Effects of DNA Phosphorothioation: B-Helical Destabilization in Rp-Phosphorothioated DNA. *The Journal of Physical Chemistry B* **116**, 10639-10648, doi:10.1021/jp302494b (2012).

228 Chen, L. *et al.* Theoretical Study on the Relationship between Rp-Phosphorothioation and Base-Step in S-DNA: Based on Energetic and Structural Analysis. *The Journal of Physical Chemistry B* **119**, 474-481, doi:10.1021/jp511359e (2015).

229 Pande, V. & Nilsson, L. Insights into structure, dynamics and hydration of locked nucleic acid (LNA) strand-based duplexes from molecular dynamics simulations. *Nucleic Acids Res.* **36**, 1508-1516, doi:10.1093/nar/gkm1182 (2008).

230 Suresh, G. & Priyakumar, U. D. Structures, Dynamics, and Stabilities of Fully Modified Locked Nucleic Acid (β-d-LNA and α-l-LNA) Duplexes in Comparison to Pure DNA and RNA Duplexes. *The Journal of Physical Chemistry B* **117**, 5556-5564, doi:10.1021/jp4016068 (2013).

231 Suresh, G. & Priyakumar, U. D. Atomistic Investigation of the Effect of Incremental Modification of Deoxyribose Sugars by Locked Nucleic Acid (β-d-LNA and α-l-LNA) Moieties on the Structures and Thermodynamics of DNA–RNA Hybrid Duplexes. *The Journal of Physical Chemistry B* **118**, 5853-5863, doi:10.1021/jp5014779 (2014).

232 Xu, Y., Villa, A. & Nilsson, L. The free energy of locking a ring: Changing a deoxyribonucleoside to a locked nucleic acid. *J. Comput. Chem.* **38**, 1147-1157, doi:10.1002/jcc.24692 (2017).

233 Pabon-Martinez, Y. V. *et al.* LNA effects on DNA binding and conformation: from single strand to duplex and triplex structures. *Sci. Rep.* **7**, 11043, doi:10.1038/s41598-017-09147-8 (2017).

234 Verona, M. D., Verdolino, V., Palazzesi, F. & Corradini, R. Focus on PNA Flexibility and RNA Binding using Molecular Dynamics and Metadynamics. *Sci. Rep.* **7**, 42799, doi:10.1038/srep42799 (2017).
235 Autiero, I., Saviano, M. & Langella, E. Molecular dynamics simulations of PNA–PNA and PNA–DNA duplexes by the use of new parameters implemented in the GROMACS package: a conformational and dynamics study. *Phys. Chem. Chem. Phys.* **16**, 1868-1874, doi:10.1039/C3CP54284J (2014).
236 Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A. D. & Nilsson, L. Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.* **37**, 896-912, doi:10.1002/jcc.24307 (2016).
237 Aduri, R. *et al.* AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA. *J. Chem. Theory Comput.* **3**, 1464-1475, doi:10.1021/ct600329w (2007).
238 Condon, D. E. *et al.* Optimization of an AMBER Force Field for the Artificial Nucleic Acid, LNA, and Benchmarking with NMR of L(CAAU). *The Journal of Physical Chemistry B* **118**, 1216-1228, doi:10.1021/jp408909t (2014).
239 Jasiński, M., Feig, M. & Trylska, J. Improved Force Fields for Peptide Nucleic Acids with Optimized Backbone Torsion Parameters. *J. Chem. Theory Comput.* **14**, 3603-3620, doi:10.1021/acs.jctc.8b00291 (2018).
240 Pokorná, P., Kruse, H., Krepl, M. & Šponer, J. QM/MM Calculations on Protein–RNA Complexes: Understanding Limitations of Classical MD Simulations and Search for Reliable Cost-Effective QM Methods. *J. Chem. Theory Comput.* **14**, 5419-5433, doi:10.1021/acs.jctc.8b00670 (2018).
241 Bergonzo, C., Henriksen, N. M., Roe, D. R. & Cheatham, T. E. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **21**, 1578-1590, doi:10.1261/rna.051102.115 (2015).
242 Jing, Z. *et al.* Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics*, doi:10.1146/annurev-biophys-070317-033349 (2019).
243 Zhang, C., Lu, C., Wang, Q., Ponder, J. W. & Ren, P. Polarizable Multipole-Based Force Field for Dimethyl and Trimethyl Phosphate. *J. Chem. Theory Comput.* **11**, 5326-5339, doi:10.1021/acs.jctc.5b00562 (2015).
244 Zhang, C., Bell, D., Harger, M. & Ren, P. Polarizable Multipole-Based Force Field for Aromatic Molecules and Nucleobases. *J. Chem. Theory Comput.* **13**, 666-678, doi:10.1021/acs.jctc.6b00918 (2017).
245 Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).
246 Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *Journal of Physical Chemistry B* **114**, 2549-2564 PMCID: PMC2918242 (2010).
247 Jing, Z., Qi, R., Liu, C. & Ren, P. Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field. *Journal of Chemical Physics* (2017).

248 Jing, Z., Liu, C., Qi, R. & Ren, P. Many-body effect determines the selectivity for Ca2+ and Mg2+ in proteins. *Proceedings of the National Academy of Sciences* **115**, E7495, doi:10.1073/pnas.1805049115 (2018).
249 Qi, R. *et al.* Elucidating the Phosphate Binding Mode of Phosphate-Binding Protein: The Critical Effect of Buffer Solution. *Journal of Physical Chemistry B* **122**, 6371-6376, doi:10.1021/acs.jpcb.8b03194 (2018).
250 Jiao, D., Golubkov, P. A., Darden, T. A. & Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *P Natl Acad Sci USA* **105**, 6290-6295, doi:10.1073/pnas.0711686105 (2008).
251 Wu, J. C., Chattree, G. & Ren, P. Automation of AMOEBA polarizable force field parameterization for small molecules. *Theor. Chem. Acc.* **131**, doi:10.1007/s00214-012-1138-6 (2012).
252 Kanaori, K. *et al.* Structure and Stability of the Consecutive Stereoregulated Chiral Phosphorothioate DNA Duplex [†]. *Biochemistry* **38**, 16058-16066, doi:10.1021/bi9909344 (1999).
253 Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible Multiple Time Scale Molecular-Dynamics. *J. Chem. Phys.* **97**, 1990-2001, doi:Doi 10.1063/1.463137 (1992).
254 Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101, doi:10.1063/1.2408420 (2007).
255 Bell, D. R. *et al.* Calculating binding free energies of host-guest systems using the AMOEBA polarizable force field. *Phys Chem Chem Phys* **18**, 30261-30269, doi:10.1039/c6cp02509a (2016).
256 Blanchet, C., Pasi, M., Zakrzewska, K. & Lavery, R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* **39**, W68-W73, doi:10.1093/nar/gkr316 (2011).
257 Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science* **271**, 795-799, doi:10.1126/science.271.5250.795 (1996).
258 Liang, X., Kuhn, H. & Frank-Kamenetskii, M. D. Monitoring Single-Stranded DNA Secondary Structure Formation by Determining the Topological State of DNA Catenanes. *Biophys. J.* **90**, 2877-2889, doi:https://doi.org/10.1529/biophysj.105.074104 (2006).
259 Markham, N. R. & Zuker, M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* **33**, W577-W581, doi:10.1093/nar/gki591 (2005).
260 Markham, N. R. & Zuker, M. in *Bioinformatics: Structure, Function and Applications* (ed Jonathan M. Keith) 3-31 (Humana Press, 2008).
261 Lesnik, E. A. & Freier, S. M. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. *Biochemistry* **34**, 10807-10815, doi:10.1021/bi00034a013 (1995).

262 Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **92**, 3817-3829, doi:10.1529/biophysj.106.097782 (2007).
263 Fink, R. F. Spin-component-scaled Møller–Plesset (SCS-MP) perturbation theory: A generalization of the MP approach with improved properties. *The Journal of Chemical Physics* **133**, 174113, doi:10.1063/1.3503041 (2010).
264 Goldey, M. B., Belzunces, B. & Head-Gordon, M. Attenuated MP2 with a Long-Range Dispersion Correction for Treating Nonbonded Interactions. *J. Chem. Theory Comput.* **11**, 4159-4168, doi:10.1021/acs.jctc.5b00509 (2015).
265 Řezáč, J., Greenwell, C. & Beran, G. J. O. Accurate Noncovalent Interactions via Dispersion-Corrected Second-Order Møller–Plesset Perturbation Theory. *J. Chem. Theory Comput.* **14**, 4711-4721, doi:10.1021/acs.jctc.8b00548 (2018).
266 Krepl, M. *et al.* Can We Execute Stable Microsecond-Scale Atomistic Simulations of Protein–RNA Complexes? *J. Chem. Theory Comput.* **11**, 1220-1243, doi:10.1021/ct5008108 (2015).
267 Kührová, P. *et al.* Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions. *J. Chem. Theory Comput.* **15**, 3288-3305, doi:10.1021/acs.jctc.8b00955 (2019).